Is the meta-analysis of correlation coefficients accurate when population correlations vary?

Article (Unspecified)

Is the Meta-Analysis of Correlation Coefficients Accurate when Population Correlations Vary?

Andy P. Field

University of Sussex, UK

Abstract

One conceptualization of meta-analysis is that studies within the meta-analysis are sampled from populations with mean effect sizes that vary (random-effects models). The consequences of not applying such models and the comparison of different methods have been hotly debated. A Monte Carlo study compared the efficacy of Hedges and Vevea's random-effects methods of meta-analysis with Hunter and Schmidt's, over a wide range of conditions, as the variability in population correlations increases. (1) The Hunter-Schmidt method produced estimates of the average correlation with the least error, although estimates from both methods were very accurate; (2) confidence intervals from Hunter and Schmidt's method were always slightly too narrow, but became more accurate than those from Hedges and Vevea's method as the number of studies included in the meta-analysis, the size of the true correlation and the variability of correlations increased and, (3) the study weights did not explain the differences between the methods.

Is the Meta-Analysis of Correlation Coefficients Accurate when Population Correlations Vary?

Meta-analysis is a statistical technique for assimilating research findings that was developed because of the failure of discursive reviews to provide objective assessments of the substantive importance of empirical effects (see Wolf, 1986). Although its objectivity can also be limited (for example, by the selective inclusion of studies and the difficulty in including all relevant studies because of unpublished research findings), Field (2001, 2003a,b) reports a remarkable increase in its usage since Glass (1976), Hedges and Olkin (1985), Rosenthal and Rubin (1978), Schmidt and Hunter (1977), Hunter, Schmidt and Jackson (1982), and Hunter and Schmidt (1990a) made their groundbreaking contributions. However, meta-analysis is not without controversy and recent debate has centered on the appropriate application of meta-analytic methods (e.g., Field, 2003a,b; Hunter & Schmidt, 2000) and comparisons of different methods (Field, 2001; Hall & Brannick, 2002; Johnson, Mullen & Salas, 1995; Schulze, 2004). This paper reviews these controversies before presenting empirical data comparing two different methods over a wide range of conditions.

Methods of Meta-Analysis

In meta-analysis, effect-size estimates from different studies are combined to try to estimate the true size of the effect in the population. Although several effect-size estimates are available (e.g., the Pearson product-moment correlation coefficient, *r*; Cohen's effect-size index, *d*; odds ratios; risk rates; and risk differences), they, at some level, represent the same thing and in some cases can be converted into one of the other metrics (see Rosenthal, 1991; Wolf, 1986)[i]. The general meta-analytic framework is similar for all of these metrics: the population effect size is estimated by

taking effect sizes for individual studies, converting them to a common metric, and then calculating a weighted average effect size that has an associated standard error. The weight given to a particular study is often based on the sample size for that study (usually the sampling variance of the effect size), which is an indicator of the sampling accuracy of that particular study. Confidence intervals can be constructed around the weighted average and its significance can be determined from a *z*-test. Meta-analysis can also be used to assess the similarity of effect sizes across studies using tests of homogeneity (Hedges & Olkin, 1985) or variance estimates (Hunter & Schmidt, 1990a; 2004).

Fixed- and Random-Effects Methods

One controversy within the meta-analysis literature is the appropriate application of methods (Field, 2003a,b; Hunter & Schmidt, 2000). In essence, there are two ways to conceptualise meta-analysis: fixed- and random-effects models (see Hedges, 1992; Hedges & Vevea, 1998; Hunter & Schmidt, 2000)[ii]. The fixed-effect conceptualisation assumes that studies in the meta-analysis are sampled from a population with a fixed effect size or one that can be predicted from a few predictors; in the simplest case, the effect size in the population is *constant* for all studies included in a meta-analysis (Hunter & Schmidt, 2000). The alternative is to assume that population effect sizes vary randomly from study to study; that is, studies in a meta-analysis come from populations of effect sizes that are likely to have different means. Population effect sizes can, therefore, be thought of as being sampled from a universe of possible effects—a 'superpopulation' (Becker, 1996; Hedges, 1992).

Which of the two conceptualisations to use is controversial and this issue hinges on both the assumptions that can realistically be made about the populations from which studies are sampled, and the types of inferences that researchers wish to make

from the meta-analysis. On the former point, there has been support for the position that real-world data are likely to have variable population parameters (Field, 2003a; Hunter & Schmidt, 1990b, 2000; National Research Council, 1992; Osburn & Callender, 1992) and empirical data have shown that real-world data do not conform to the assumption of fixed population parameters (Barrick & Mount, 1991). Figure 1 shows the distribution of between-studies standard-deviation estimates calculated (where data were available) for all meta-analytic studies using correlation coefficients published in *Psychological Bulletin* 1997-2002[iii]. These estimates are based on the Hunter-Schmidt method (they are the square root of equation (15), described later). This histogram shows that the meta-analytic studies typically give rise to between-studies standard deviation estimates ranging from 0 to 0.3, with values of 0 being very common. However, when effect-size variability is present, it is most frequently in the region of 0.10-0.16 (which is broadly consistent with Barrick & Mount, 1991), and values as high as 0.3 are relatively infrequent.

With regard to the latter point, Hedges and Vevea (1998) suggested that the choice of model depends on the type of inferences that the researcher wishes to make: fixed-effect models are appropriate for inferences that extend only to the studies included in the meta-analysis (*conditional inferences*) whereas random-effects models allow inferences that generalise beyond the studies included in the meta-analysis (*unconditional inferences*). Psychologists typically wish to make generalizations beyond the studies included in the meta-analysis and so random-effects models are more appropriate (Field, 2003a; Hunter & Schmidt, 2000).

This debate has been exacerbated by the fact that fixed- and random-effects meta-analytic methods are frequently incorrectly applied. Despite some good evidence that real-world data support a random-effects conceptualisation, psychologists routinely apply fixed-effects meta-analytic methods to their data. For example, Hunter

and Schmidt (2000) listed 21 recent examples of meta-analytic studies using fixed-effects methods in the major review journal for psychology (*Psychological Bulletin*) compared to none using random-effects models. The theoretical consequence, according to Hunter and Schmidt (2000), is that the significance tests of the average effect size should not control the Type I error rate: they predicted inflated error rates of between 11–28%. In fact, Field (2003a) has shown using Monte Carlo simulations that Type I error rates are inflated from 5% to anywhere between 43 and 80%. So, of the 21 meta-analyses reported by Hunter and Schmidt (2000) anywhere between 9 and 17 of them are likely to have reported significant effects when in reality no true effect may have existed within the population (see Field, 2003a).

One function of the homogeneity of effect-size measures mentioned earlier is to ascertain whether population effect sizes are likely to be fixed or variable, through inference from the variability in sample effect sizes (Hedges & Olkin, 1985). The rationale is that if homogeneity tests yield non-significant results then sample effect sizes are roughly equivalent and so population effect sizes are likely to be homogenous (and hence the assumption that they are fixed is reasonable). Even if we overlook the fact that researchers often ignore these tests (e.g., 17 of the 21 meta-analyses listed by Hunter & Schmidt, 2000, used fixed-effect methods despite significant homogeneity tests), the tests themselves can be misleading because they sometimes have been claimed to have low power to detect genuine variation in population effect sizes (Hedges & Pigott, 2001; Sackett, Harris & Orr, 1986; but see Field, 2001, who showed that their power is high). Consequently, researchers can be misled into concluding that population effect sizes are fixed when they are, in fact, variable.

Comparing Methods

Since the early work of Glass (1976) two methods of meta-analysis have remained popular: the methods devised by Hedges and colleagues, and those of Hunter and Schmidt (1990a, 2004)[iv]. Hedges and colleagues (Hedges & Olkin, 1985; Hedges, 1992; Hedges & Vevea, 1998) have developed both fixed- and random-effects models for combining effect sizes, whereas Hunter and Schmidt (and Hunter, Schmidt & Jackson, 1982) label their method a random-effects model (see Hunter & Schmidt, 2004; National Research Council, 1992; Schmidt & Hunter, 1999) although in earlier writings they were less explicit in defining it in these terms (Hunter & Schmidt, 1990a). The second main controversy in the meta-analysis literature is which of these two methods should be applied. This paper looks at methods for combining effect sizes expressed as correlation coefficients, *r*.

Hedges and Colleagues' Method

In this method, correlations are first converted into a standard normal metric (using Fisher's *r*-to-*Z* transformation) before calculating a weighted average of these transformed scores. Fisher's (1921) *r*-to-*Z* transformation is given in equation (1) in which r_i is the correlation coefficient from study *i*

$$z_{r_i} = \frac{1}{2} Log_e \left(\frac{1 + r_i}{1 - r_i} \right), \qquad (1)$$

which has an approximate normal distribution with mean $\bar{z}_\rho$, and variance $1/(n_i - 3)$, where n_i is the number of cases or pairs of data in the study. The transformation back to r_i is simply

$$r_i = \frac{e^{(2z_i)} - 1}{e^{(2z_i)} + 1} . \quad (2)$$

The transformed effect sizes are then used to calculate an initial average in which each correlation is weighted by the inverse of the within-study variance of the study from which it came (for Fisher z_r values the sample size, n_i, minus three)— see Equation (3)[v], and where k is the number of studies in the meta-analysis (Hedges & Olkin, 1985, p. 231):

$$\bar{z}_r = \frac{\sum_{i=1}^{k} w_i z_{r_i}}{\sum_{i=1}^{k} w_i} = \frac{\sum_{i=1}^{k} (n_i - 3) z_{r_i}}{\sum_{i=1}^{k} (n_i - 3)} . \quad (3)$$

This average is then used to calculate a test of the homogeneity of correlations: the squared difference between each study's observed transformed r and the mean transformed r (from equation (3)), weighted by the within-study variance, is used. This gives us the statistic Q in Equation (4), which has a chi-square distribution with k – 1 degrees of freedom under the null hypothesis of homogenous effect sizes (Hedges & Olkin, equation 16, p. 235):

$$Q = \sum_{i=1}^{k} (n_i - 3)(z_{r_i} - \bar{z}_r)^2 . \quad (4)$$

To calculate the random-effects average correlation, the weights use a variance component that incorporates both between-studies variance and within-study variance. The between-studies variance is denoted by τ^2 and an estimate of it ($\hat{\tau}^2$) is simply added to the within-study variance. The weighted average in the z_r metric is (based on Hedges & Vevea, 1998, equation 12):

$$\bar{z}_r^* = \frac{\sum_{i=1}^{k} w_i^* z_{r_i}}{\sum_{i=1}^{k} w_i^*}, \quad (5)$$

in which the weights (w_i^*) are defined as (based on Hedges & Vevea, 1998, equation 14):

$$w_i^* = \left(\frac{1}{n_i - 3} + \hat{\tau}^2\right)^{-1}. \quad (6)$$

The between-studies variance can be estimated in several ways (see Friedman, 2000; Hedges & Vevea, 1998; Overton, 1998; Takkouche, Cadarso-Suárez & Spiegelman, 1999), however, Hedges and Vevea (1998, equation 10) use Equation (7)), which is based on *Q* (the weighted sum of squared errors in equation (4)), *k*, and a constant, *c*, such that:

$$\hat{\tau}^2 = \frac{Q - (k-1)}{c}, \quad (7)$$

where the constant, *c*, is defined as:

$$c = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} (w_i)^2}{\sum_{i=1}^{k} w_i}, \quad (8)$$

and for correlations, because $w_i = n_i - 3$, *c* is:

$$c = \sum_{i=1}^{k} (n_i - 3) - \frac{\sum_{i=1}^{k} (n_i - 3)^2}{\sum_{i=1}^{k} (n_i - 3)}. \quad (9)$$

If the estimate of between-studies variance, $\hat{\tau}^2$, yields a negative value then it is set to zero (because the variance between-studies cannot be negative). The estimate $\hat{\tau}^2$, is substituted in equation (6) to calculate the weight for a particular study, and

this in turn is used in equation (5) to calculate the average correlation. This average correlation is then converted back to the *r* metric using equation (2) before being reported.

The sampling variance of the untransformed average correlation is the reciprocal of the sum of weights and the standard error of this average correlation is the square root of this sampling variance. Bearing in mind that the weights are calculated using equation (6) the standard error is (see Hedges & Vevea, 1998, p. 493):

$$SE\left(\overline{z}_r^*\right) = \sqrt{\frac{1}{\sum_{i=1}^{k} w_i^*}}\ . \qquad (10)$$

Hedges and Olkin (1985) recommend constructing a confidence interval around the average effect size, which is easily done using the standard error and $z_{\alpha/2}$, the two-tailed critical value of the normal distribution (which is 1.96 for the most commonly used 95% confidence interval). The upper and lower bounds are calculated by taking the average effect size from equation (5) and adding or subtracting its standard error multiplied by 1.96:

$$\begin{aligned} CI_{Upper} &= \overline{z}_r^* + 1.96SE\left(\overline{Z}_r^*\right), \\ CI_{Lower} &= \overline{z}_r^* - 1.96SE\left(\overline{Z}_r^*\right). \end{aligned} \qquad (11)$$

These values are again transformed back to the *r* metric using equation (2) before being reported.

Hunter and Schmidt Method

This method emphasises the need to isolate and correct for sources of error such as sampling error and reliability of measurement scales. Although these recommended corrections are undoubtedly the method's great strength, this study deals only with

the method in its simplest form. Hunter and Schmidt (2004, p. 81) recommend using untransformed effect-size estimates, r_i, to calculate the weighted mean correlation, and the weight used is simply the sample size, n_i:

$$\bar{r} = \frac{\sum_{i=1}^{k} n_i r_i}{\sum_{i=1}^{k} n_i} . \qquad (12)$$

Hunter and Schmidt (2004) argue that the variance across sample correlations will be made up of the variance of correlations in the population and the sampling error; therefore, to estimate the variance in population correlations we have to correct the variance in sample correlations by the sampling error. The variance of sample correlations is the frequency weighted average squared error. Equation (13), from Hunter and Schmidt (2004, p. 81 & 89), shows this:

$$\hat{\sigma}_r^2 = \frac{\sum_{i=1}^{k} n_i (r_i - \bar{r})^2}{\sum_{i=1}^{k} n_i} . \qquad (13)$$

The sampling error variance is calculated using the average correlation, $\bar{r}$, and the average sample size, $\overline{N}$, (see Hunter & Schmidt, 2004, p. 88):

$$\hat{\sigma}_e^2 = \frac{\left(1-\bar{r}^2\right)^2}{\overline{N}-1} . \qquad (14)$$

It is a simple matter to estimate the variance in population correlations by subtracting the sampling error variance from the variance in sample correlations (see Hunter & Schmidt, 2004, p. 88):

$$\hat{\sigma}_{\rho}^2 = \hat{\sigma}_r^2 - \hat{\sigma}_e^2 . \quad (15)$$

Hunter and Schmidt recommend correcting this estimate for artifacts (see Hunter & Schmidt, 2004 or Hall & Brannick, 2002 for details) and then constructing what they call credibility intervals. These intervals are based on taking the average correlation (equation (12)) and adding to or subtracting from it the square root of the estimated population variance in equation (15) multiplied by $z_{\alpha/2}$ (1.96 for a 95% interval):

$$\begin{aligned} \text{Credibility Interval}_{\text{Upper}} &= \bar{r} + 1.96\sqrt{\hat{\sigma}_{\rho}^2} , \\ \text{Credibility Interval}_{\text{Lower}} &= \bar{r} - 1.96\sqrt{\hat{\sigma}_{\rho}^2} . \end{aligned} \quad (16)$$

If confidence intervals are required (rather than credibility intervals) these can be obtained by using the standard error of the mean correlation. To obtain this standard error simply divide the variance of sample correlations (given in equation (13)) by the number of studies in the meta-analysis, *k*, and take the square root:

$$\begin{aligned} CI_{Upper} &= \bar{r} + 1.96\sqrt{\frac{\sigma_r^2}{k}} , \\ CI_{Lower} &= \bar{r} - 1.96\sqrt{\frac{\sigma_r^2}{k}} . \end{aligned} \quad (17)$$

Differences between Methods

If we take the bare bones version of the Hunter-Schmidt method (described above), which does not correct for research artifacts other than sampling error, then the differences between this method and Hedges' random-effects method are: (1) the use of transformed or untransformed correlation coefficients, and (2) the difference in study weighting (which in turn creates differences in the estimates of the sampling error variance of the mean).

The evidence on whether or not it is better to transform *r* has been inconsistent. For example, although Silver and Dunlap (1987) claimed that meta-analysis based on Fisher transformed correlations is always less biased than when untransformed correlations are used, they ignored the effect of the number of studies in the analysis: they did not report results for when different numbers of correlation coefficients were being combined, nor did they report how many correlation coefficients were combined for the data presented. Strube (1988) went on to demonstrate that as the number of studies increased there was no discernible difference between the biases resulting from using Fisher transformed or untransformed correlations. In addition, transformed correlations were less biased than untransformed ones only when 3 or fewer studies were included in the meta-analysis and sample sizes were 20 or less (which would be rare in the application of meta-analysis). However, this study too was limited to the scenario in which a maximum of 20 studies were included in the meta-analysis. Schulze (2004) in a set of extensive simulations recently concluded that computations based on *z* transformed values would invariably differ from those based on untransformed *r*s.

The second difference is in the study weighting[vi]. Hedges and Vevea (1998) have argued that Hunter and Schmidt's method assumes that the between-studies variance is small; therefore, when between-studies variance is not small, the practice of weighting studies by n_i (see equation 12) in the Hunter-Schmidt method should produce an inaccurate mean correlation (Hedges & Vevea, 1998). However, although both methods have estimates of the between-studies variance that are truncated at zero (if computed values are negative, they are set to zero), unlike the H-S method, Hedges and Vevea's (1998) method uses the estimate of between-studies variance as part of the study weights. Consequently, the accuracy of the average correlation is

biased in this method too — especially when the number of studies in the meta-analysis is small (see Hedges & Vevea, 1998).

Several recent studies have attempted to compare these two methods. Johnson et al. (1995) compared the Hedges-Olkin (fixed-effect), Rosenthal-Rubin and Hunter-Schmidt meta-analytic methods by manipulating a single data set to look at the effects of the number of studies compared, the mean correlation of studies, the mean number of participants per study and the range of effect sizes within the database. They concluded that the methods converged in terms of the mean correlation and estimates of the heterogeneity of effect sizes, but the significance of the mean correlation differed substantially across the methods: the Hunter and Schmidt method reached more conservative estimates of significance than the other two methods and Johnson et al. concluded that it should be used only with caution.

Although this study was a good starting point, Schmidt and Hunter (1999) claimed that Johnson et al. used the wrong estimate of the standard error of the mean correlation and showed that, theoretically, when a corrected estimate was used, their method was comparable to the Hedges and Olkin and Rosenthal and Rubin methods. Field (2001) also pointed out that Johnson et al. applied Hedges and Olkin's method for *d* (by first converting each correlation coefficient from *r* to *d*) rather than using the methods for directly combining *r*s (without converting to *d*), that the use of a single database limited the generality of the findings, and that the Hunter-Schmidt method had not been compared with Hedges' *random*-effects counterpart. Field (2001) rectified some of these concerns by conducting a series of Monte Carlo simulations comparing the performance of the Hunter and Schmidt and Hedges and Olkin (fixed- and random-effects) methods both when population correlations were fixed, and when they were variable. The number of studies in the meta-analysis, the average sample size, and the size of the correlation in the population were systematically varied. Field

found that when comparing random-effects methods the Hunter-Schmidt method yielded the most accurate estimates of population correlation across a variety of situations. However, neither the Hunter-Schmidt nor Hedges and colleagues' method controlled the Type I error rate when 15 or fewer studies were included in the meta-analysis, and the method described by Hedges and Vevea (1998) controlled the Type I error rate better than the Hunter-Schmidt method when 20 or more studies were included. Hall and Brannick (2002) conducted a similar study but looking at the methods within the context of test validation and found that the Hunter and Schmidt method provided the most accurate estimates of the population correlation. Schulze (2004) has also done extensive simulation studies and based on these findings recommends against using Fisher's *z* transform and suggests that the 'optimal' study weights used in the H-V method can, at times, be sub-optimal in practice. However, Schulze based these conclusions on using the fixed-effects version of Hedge's method (he did not examine the method described by Hedges and Vevea, 1998)

The Current Study

Although Field (2001) and Hall and Brannick (2002) have used simulation techniques to compare these two methods under a wider variety of situations than earlier researchers, their findings are still limited. Field restricted his simulations to a single degree of variability between effect sizes and extended his simulations only to when meta-analyses of 30 studies were included (relatively few in real terms). Hall and Brannick used a fairly restricted range of population correlation variances. Furthermore, neither study systematically investigated the source of the differences between methods. Schulze (2004) took a different approach in his simulations and used discrete distributions of the true correlation and it is unclear how much the simulation process itself contributes to the conclusions drawn about the relative merits

of the two methods. He also did not explicitly investigate Hedges and Vevea's (1998) random-effects method. As such, the current study aims to extend Field's (2001), Hall and Brannick's (2002) and Schulze's (2004) work by (1) comparing two random-effects methods across a more diverse array of situations than Field and Hall and Brannick, (2) investigating whether technical aspects of the simulations affect the conclusions drawn, and (3) investigating whether the weights in the two methods are responsible for the differences observed. A few general predictions can be made:

(1) Confidence intervals and estimates of the mean correlation from the H-V method should be less accurate when small numbers of studies are included in the meta-analysis and variability between effect sizes is small. This is because the study weights are based on estimates of between-studies variance that are truncated at zero.

(2) If Hunter and Schmidt's method utilizes sub optimal weights that do not take account of the variability between population effect sizes (Hedges & Vevea, 1998), then estimates of the average correlation and its variability will become less accurate as the variability between population effect sizes increases.

(3) A related prediction is that if the weights are responsible for the differences between the methods, the estimates from the H-S method should improve if the H-V 'optimal' weights are used (conversely, the H-V method should become less accurate if the weights are replaced with those from the H-S method).

(4) If the z-transformation is a useful procedure then the Hunter-Schmidt method should become less accurate (especially for small samples) as the correlation and standard deviation in the superpopulation increases (because as these parameters increase, the resulting population distribution becomes non-normal: negatively skewed and platykurtic). Conversely, Hedges' method, because it

uses Fisher's transformation, should become relatively more accurate when the population distributions are skewed. However, if, as Schulze (2004) suggests, the Fisher's *z*-transformation is unnecessary then the skew of the distribution of *r*s will not adversely affect H-S estimates.

(5) Simulating meta-analysis using a superpopulation of *z*-transformed values should favour the H-V method whereas a superpopulation based on *r* should favour the H-S method.

Monte Carlo Simulations

General Method

The rationale behind all of the simulations in this paper is simple: if data are sampled from a population which itself is sampled from a superpopulation with a known mean and standard deviation, the accuracy of random-effects methods can be ascertained by comparing the mean correlation against the known mean in the superpopulation. The standard deviation within the superpopulation can also be manipulated to look at how variability between population effect sizes influences the accuracy of different meta-analytic methods (in terms of their average correlation and the confidence intervals around that average).

The general approach was as follows: a distribution of correlations with a known average and standard deviation was created to act as a superpopulation, from which the population correlation for each study in a meta-analysis was sampled. A sample of a given size was taken from a particular population and the correlation coefficient calculated and stored. Once a specified number of samples (representing the number of studies in the meta-analysis) had been taken from the populations, the two random-effects meta-analytic techniques were applied (the average correlation and 95% confidence intervals were calculated). The number of samples in the meta-

analysis, the relative size of those samples, and the standard deviation of superpopulation correlations were varied systematically to look at whether these factors influenced the accuracy of the methods. Also, by looking at the average correlation across many trials it was possible to ascertain the boundaries within which 95% of average correlations fell. As such, a *population* confidence interval for the average correlation could be calculated by discarding the lower and upper 2.5% of estimated average correlations from 100,000 trials.

The Superpopulation

All simulations were run using GAUSS 4.0. A distribution of possible correlations was created (a superpopulation) from which the population correlation for each study in a meta-analysis was sampled. Although this distribution was theoretically normal, the exact shape of the distribution of ρs sampled from this distribution depended on the values of the mean and standard deviation. This is discussed in due course (and in Figure 3). The mean effect size of the superpopulation ($\overline{\rho}$) was initially set to be 0 and was then systematically changed to represent a small ($\overline{\rho}$ = .1), medium ($\overline{\rho}$ = .3), large ($\overline{\rho}$ = .5), and a huge effect size ($\overline{\rho}$ = .8) based on Cohen's (1988, 1992) guidelines for correlation coefficients.

As the mean correlations and their standard deviations get larger, values in the superpopulation will begin to exceed 1 (the upper bound of the correlation coefficient). To prevent this, two methods were used to define this superpopulation:

1. The superpopulation was treated as a distribution of *z*-transformed correlations (see equation (1)), which are not constrained to be less than 1. So, the z-transformation was first used to transform the mean correlation (0, .1, .3, .5 or .8), then a normal distribution with this mean and a specified standard deviation (0.04, 0.08, 0.16 or 0.32)[vii] was created. A correlation was sampled from this

distribution and was back-transformed (see equation (1)) to create the correlation in the population (which would be less than 1).

2. The superpopulation was treated as a distribution of *ρ*s, but with a correction to prevent inadmissible values. The mean correlation (0, .1, .3, .5 or .8) and standard deviation (0.04, 0.08, 0.16 or 0.32) of correlation was set as before, and then a normal distribution with this mean was created. A correlation was sampled from this distribution but if its value was greater than 1 or less than -1, it was rejected and a new correlation was sampled.

Once the mean correlation for the population had been sampled from the superpopulation, it was necessary to simulate the process of 'sampling' the sample correlation from this population. This was achieved using the **A** matrix procedure described by Mooney (1997) in which the correlation between two randomly generated normally distributed variables is set using the Choleski decomposition of a fixed correlation matrix. The correlation matrix contained a single value, ρ, and pairs of normally distributed scores for n_i cases were then generated (n_i was manipulated as described below) using this value and the **A** matrix procedure already described. The correlation between these two variables was then calculated and represented the sample correlation.

For each Monte Carlo trial a set number of studies was taken from a given population and the average correlation and its confidence interval were calculated using both methods. The proportion of these confidence intervals containing the effect size in the superpopulation was calculated over 100,000 Monte Carlo trials, and *population* confidence intervals were constructed (defined earlier).

The Mean and Standard Deviation of Population Correlations

When the superpopulation was treated as a distribution of *ρ*s, the mean correlation of the superpopulation ($\bar{\rho}$) was initially set to be 0 and was then systematically changed to represent a small ($\bar{\rho}$ = .1), medium ($\bar{\rho}$ = .3), large ($\bar{\rho}$ = .5), and a huge correlation ($\bar{\rho}$ = .8). The standard deviation of the superpopulation was systematically varied to be 0.04, 0.08, 0.16, and 0.32. These were set based on values of the estimates of standard deviation of population correlations reported in *Psychological Bulletin* during 1997-2002 (see Figure 1). The end result is a set of population standard deviations that represent situations that range from almost no variability (0.04) through frequently reported variability (0.04 to 0.16) to around the largest reported variability (0.32). The intention was to present a full range of situations so that data can fully inform real-world situations including values at the very extreme of what could be found (i.e., 0.32).

When the superpopulation was treated as a distribution of *z*-transformed correlations, the same values for the means and standard deviations were used but were applied to the distribution of *zs*; therefore, the actual means and standard deviations of the back-transformed distribution (the distribution of *ρ*s) were smaller. Values of the actual means and standard deviations for all combinations used in the main study were estimated using the data simulation function of Excel with 10000 data points generated for each cell. The resulting values can be found in Table 1.

Number of Studies

The number of studies, *k*, used in the meta-analysis was manipulated by varying it from 5 to 160, with 6 values: 5, 10, 20, 40, 80 and 160. Field (2001) varied the number of studies in the meta-analysis only from 5 to 30, so the range of numbers of studies in the present simulation is much wider.

Average Sample Size

The average size of each sample in the meta-analysis was also varied. In most real-life meta-analyses study sample sizes are not equal; so to model reality sample sizes were drawn from a normal distribution of possible sample sizes, with the mean of this distribution being systematically varied. Sample sizes were randomly taken from a distribution with a fixed mean (20, 40, 80 or 160) and a standard deviation of a quarter of that mean. Although by taking sample size values from a population an additional source of variance is introduced into the simulations, this approach was taken because in real-life meta-analyses study sample sizes do vary (otherwise there would be no point in weighting studies based on their sampling accuracy). As such this method more accurately represents what happens in reality than if sample sizes were held constant (Field, 2003a, 2001 and Hall & Brannick, 2002 used similar methods).

Values of the average sample size, $\overline{n}$, were set using estimates of the sample size necessary to detect small, medium and large effects in the population based on a single study. The sample sizes needed to detect a small, medium or large effect in a single study with power = .8 are approximately 150, 50, 25 respectively (see Cohen, 1988). As such, values were set at $\overline{n}$ = 20, 40, 80 and 160. The one restriction was that any sample sizes less than 4 were discarded and replaced because Hedges' method of meta-analysis requires a sample size of at least this size.

Design

In all simulations the design was a four factor, 5 (mean superpopulation correlation: 0, .1, .3, .5, .8) × 4 (standard deviation of superpopulation: 0.04, 0.08, 0.16, 0.32) × 4 (average sample size: 20, 40, 80, 160) × 6 (number of studies: 5, 10, 20, 40, 80, 160), design. For each level of these 480 combinations 100,000 Monte

Carlo trials were used (100 times as many as the minimum recommended by Mooney, 1997). Each cell of the design contained 100,000 cases of data (48,000,000 samples of data were simulated in all).

Having said this, the results for certain combinations of the mean superpopulation correlation and standard deviations were omitted because they represent unrealistic situations. In reality, values of ρ in the superpopulation would never reach their maximum value of 1 because of the measurement error that is always present in whatever measures are used. Even assuming extremely high reliability of measures (e.g., .9), and the maximum possible correlation between constructs of 1, the maximum value of ρ in the superpopulation would be:

$$[.9(.9)]^{1/2} \times 1 = .9.$$

As such, values above .9 in the superpopulation are unrealistic[viii]. By converting .9 to a *z*-score and using tables of the normal distribution (e.g., Field, 2005) it is clear, for example, that in a superpopulation with a mean correlation of .8 and a standard deviation of 0.32 that (z = (.9-.8)/0.32 = 0.31) 37.83% of ρs in the superpopulation will be above .9. Results are reported for situations in which less than 5% of ρs in the superpopulation fall above .9; so, all levels of the standard deviation are reported for mean correlations up to .3 (maximum number of ρs above .9 = 3.04%), but only standard deviation values up to 0.16 are reported for a superpopulation mean of .5 and only a standard deviation value of 0.04 is reported for the maximum superpopulation mean of .8 (in both cases only 0.62% of ρs would fall above .9).

Results

Simulation 1: Comparing the Methods

The first simulation compared the average correlations and their confidence intervals when the superpopulation was constructed using *z*-transformed values (described as method 1 above). Figure 2 shows the estimated average correlation and the boundaries between which 95% of average correlations from the Monte Carlo simulations fell; both are expressed in terms of the deviation from the true correlation (values of which are in Table 1) and so can be directly compared across the different parameters of the simulation.

Hedges' method (H-V) yielded average correlations within .031 of the actual correlation in all conditions (the maximum deviation being when the superpopulation correlation was .3 and the standard deviation of the superpopulation was large (0.32)—in all other conditions average correlations were within .015 of the actual correlation. These results are inconsistent with Field (2001) who found that the H-V method substantially overestimated the mean correlation. The Hunter-Schmidt (H-S) method generally produced more accurate average correlations than the H-V method with the maximum deviation being −.010 (which occurred when the actual correlation was at the extreme value of .80). Although these findings are in line with Hedges and Vevea's (1998) belief that because Hunter and Schmidt's method does not weight studies using the between study variance it will underestimate the average correlation, these underestimations were minimal and less than the overestimation produced by the H-V method. When the mean superpopulation correlation was zero the H-S estimates were always accurate, when the superpopulation correlation increased to .1, .3, .5, and .8, the underestimation ranged from .000 to .003, .000 to .007, .000 to .010, and .001 to .010 respectively.

Figure 2 also shows the population confidence intervals for the two methods: these are empirically determined limits within which 95% of observed population effect size estimates fell. These population confidence intervals were fairly comparable across methods. Generally, the confidence intervals became tighter around the average as the number of studies in the meta-analysis increased but became wider as the standard deviation of superpopulation correlations increased. For example, when there was no effect in the superpopulation, average correlations ranged from, in their extreme, -.35 to +.35 when the standard deviation of superpopulation correlations was extreme (σ_ρ = 0.32) and the number of studies in the meta-analysis was small (*k* = 5). For both methods, when the standard deviation of superpopulation correlations was large (σ_ρ = 0.16 to 0.32) it was possible to obtain erroneously small to medium average correlations unless about 40 (σ_ρ = 0.16) or 80 (σ_ρ = 0.32) studies were included in the meta-analysis. However, when the standard deviation of superpopulation correlations was small (σ_ρ = 0.04 to 0.08) small to medium average correlations were obtained only when 20 or fewer studies were included in the meta-analysis. When there was a non-zero effect in the superpopulation, to keep average correlation estimates within .1 of the true value, then 40 or more studies needed to be included if the mean effect in the superpopulation was small to medium ($\overline{\rho}$ = .1 or .3), and the standard deviation of correlations, σ_ρ, was greater than or equal to 0.16. If the correlation in the superpopulation was large ($\overline{\rho}$ = .5), or if the standard deviation of correlations was smaller (σ_ρ = 0.08) only 10-20 or more studies were required to keep estimates within .1 of the true value. However, when the standard deviation of superpopulation correlations was small (σ_ρ = 0.04) population confidence intervals were fairly tight even with only 5 studies in the meta-analysis. There was also a relationship between the average sample size of studies and the standard deviation of effect sizes: at large standard deviations (σ_ρ = 0.32) the confidence

intervals were relatively unaffected by differences in study sample sizes, but as the standard deviation of effect sizes got smaller, the confidence intervals became wider as study sample sizes decreased.

Simulation 2: Does the Shape of the Superpopulation Influence the Results?

As the mean correlation in the superpopulation increases and as the standard deviation of this population increases, the distribution of correlations becomes skewed because correlation coefficients cannot exceed 1, and so there will be a build up of sample correlations just below 1. One way to model this is as in the first simulation: the superpopulation was treated as a population of *z*-transformed correlations (see equation (1)), which are not constrained to be less than 1. For the H-S method the sampled correlations were back-transformed to a correlation coefficient (see equation (1)) before any calculations were carried out. Regardless of the relative performance of the H-S and H-V methods, this simulation method has the potential to make the H-V estimates better (because they are based on *z*-transformed correlation coefficients), and the H-S estimates worse, than they actually are. Simulation 2 sought to test this possibility by re-running simulation 1, but using a different superpopulation. In this simulation, the superpopulation was made up of *r* values (as described in method 2 above), but inadmissible correlations were rejected and a new correlation sampled until one was found with a value between -1 and 1.

The effect of these two methods of simulation can be seen in Figure 3[ix], which shows the frequency distribution of correlations in the population (when the superpopulation is based on z_r values, these values have been transformed back into *r* for these graphs) as the mean correlation and its associated standard deviation in the superpopulation changes. The main difference is what happens as the mean correlation in the superpopulation becomes large or huge ($\bar{\rho}$ = .5 or greater): when

the superpopulation contains *z*-transformed values, the distribution becomes leptokurtic, whereas when raw *r*-values are used, the distribution retains a shape similar to that for smaller effects. As the standard deviation of correlations increases to 0.16 or more, the distributions start to deviate from normal distributions (not just for large mean correlations). Table 2 shows the values of skew and kurtosis for each of these distributions and whether the associated *z*-test is significant[x].

Another thing to note from Figure 3 is that when the superpopulation is based on *r* (as in the current simulation), the distributions become slightly truncated. One consequence of this truncation is that the true mean correlation for these populations will not be the value set in the simulations (i.e., the true effects will not be 0, .1, .3, .5 and .8). Therefore, to evaluate the accuracy of the estimates from the H-V and H-S methods of meta-analysis it is important we compare these estimates to the truncated mean of the superpopulation and *not* the value set in the simulation. Schmidt, Hunter & Urry (1976) present equations for calculating the mean and standard deviation of a truncated normal distribution. For this example (using correlations truncated at the top end of the distribution) the true mean of the distribution is given by equation (18) in which pdfn(x) is the Normal probability density function (pdf) of x, cdfn(x) is the cumulative distribution function (cdf) of the Normal distribution at x, and $\overline{\rho}$ and σ_ρ are the mean and standard deviation respectively of the distribution before truncation. Throughout this simulation, mean correlations from the H-V and H-S methods were compared to the true effect in the superpopulation based on equation (18).

$$\bar{\rho}_{Truncated} = \left(\frac{\text{pdfn}\left(\frac{1 - \bar{\rho}}{\sigma_\rho} \right)}{1 - \text{cdfn}\left(\frac{1 - \bar{\rho}}{\sigma_\rho} \right)} \times \sigma_\rho \right) + \bar{\rho} \qquad (18)$$

Figure 4 shows the estimates of the average correlation and the boundaries between which 95% of average correlations from the Monte Carlo simulations fell expressed as deviations from the true effect in the superpopulation. For the H-V method, the results were, as predicted, worse than those of simulation 1: the *profile* of results was similar to simulation 1 in that this method still tended to over-estimate the true correlation, but these overestimations were larger than simulation 1. As the mean superpopulation correlation (before truncation) increased from 0, .1, .3, .5 to .8 the overestimations ranged from .000 to .002, .000 to .018, .001 to .052, .002 to .031, and .004 to .012 respectively. The H-S method performed relatively similarly under these simulation conditions to those in simulation 1. As the mean superpopulation correlation (before truncation) increased from 0, .1, .3, .5 to .8 the deviations ranged from .000 to .002, .000 to –.003, .000 to –.007, –.001 to –.011, and .000 to –.008 respectively.

Confidence Intervals from Simulations 1 and 2

Table 3 shows the proportion of confidence intervals calculated using the H-V equations that contained the true correlation from the superpopulation (the values in Table 1). These proportions ranged from .814 to .962. When the standard deviation of correlations was small (σ_ρ = 0.04) the proportion of confidence intervals containing the true effect sizes was between .83 and .96. The vast majority of proportions fall between .94 and .96 (within .01 of the desired .95), the only exceptions were when the number of studies combined was large (k = 160) and the true correlation was .3

or .5, and when 40 or more studies were combined and the true correlation was .80. As the standard deviation of correlations increased to 0.08, more confidence intervals failed to include the true correlation. When the true correlation was small ($\bar{\rho}$ = 0 or .1), most proportions were between .94 and .96, however, for larger population correlations proportions fell between .94 and .96 only when average sample sizes were less than 40 and fewer than 80 studies were combined. When the standard deviation of correlations was 0.16, confidence intervals included the true correlation on 94-96% of occasions only when the true correlation was small ($\bar{\rho}$ = 0 or .1) and 40 or more studies were combined. In all other situations less than 94% of confidence intervals contained the true correlation. Finally, when the standard deviation of correlations was large (σ_ρ = 0.32), fewer than 94% of confidence intervals contained the true correlation except when the true effect was zero and more than 40 studies were combined.

Table 4 shows the proportion of confidence intervals calculated using equations based on H-S's methodology that contain the true correlation from the superpopulation (based on Table 1). It is worth remembering that H-S advocate the use of artifact-corrected credibility intervals, and not the confidence intervals constructed here (which have been used because they are comparable to the H-V confidence intervals). Nevertheless, the proportions for the H-S model were lower than the H-V model in general (they ranged from .830 to .949). The proportions were above .94 only when 80 or more studies were included in the meta-analysis and when the true correlation was .3 or less. When the true correlation was .3 or less, the proportions got closer to the desired .95 as the number of studies in the meta-analysis increased; however, when the true correlation was .5 or larger this was true only up to 40 studies included in the meta-analysis, with proportions dropping away from .95 as more studies were included. There was no condition for which the

proportion reached the desired .95 suggesting that confidence intervals were too narrow.

Tables 5 (H-V) and 6 (H-S) show the same information as Tables 3 and 4 but when the superpopulation was based on values of *r* (in these cases the true correlations are those calculated using equation (18)). For the H-V method, the results are broadly similar using this different superpopulation. The main differences are that proportions are lower using a superpopulation based on *r* when the true correlation was .5 or more, 80 or more studies were in the meta-analysis and the standard deviation of correlations was 0.08 or less. When the standard deviation of correlations was 0.16, the proportions were lower using a superpopulation based on *r* in all conditions when the true correlation was .3 or larger. When the standard deviation of correlations was 0.32, the proportions were lower using a superpopulation based on *r* when the true correlation was .1 or larger and 80 or more studies were included in the meta-analysis. For the H-S method the results when the simulation was based on a superpopulation of *r* values were virtually identical to when the superpopulation was based on *z*-transformed values. As such, the method of simulating the superpopulation made a slight difference for the H-V method (which performed slightly better when the superpopulation contained *z*-transformed values), but made very little difference in how well the H-S method performed.

To sum up, coverage proportions for H-S's method were always too low: these confidence intervals did not capture the true effect in the superpopulation as often as they should. Coverage proportions from the H-V method were generally on target; however, when the variability of effect sizes was large the coverages were sometimes much lower than for H-S.

Simulations 3 and 4: The Effect of the Weights

One suggestion has been that the H-S method uses sub-optimal weights and so is inaccurate in random-effects situations. Although the results shown in Figures 2 and 4 do not support this suggestion, simulations 3 and 4 tested this possibility in another way by replicating simulations 1 and 2, but calculating each method using the weights from the opposite method. For the H-V method this entailed replacing the w_i^* in equations (5) and (10), with n_i. For the H-S method, this entailed replacing n_i with w_i^* in equations (12) and (13). This process was used to try to tease apart the effect (if any) that the weights have on the accuracy of estimates of the mean correlation—*it is not recommended for typical practice*.

Figure 5 shows the estimates of the average correlation and the boundaries between which 95% of average correlations from the Monte Carlo simulations fell expressed as deviations from the true size of the correlation. The results were virtually identical to simulation 1 (compare Figures 2 and 5): the change in weights had virtually no effect on the estimates of the true correlation.

Table 7 shows the proportion of confidence intervals calculated using the H-V equations that contain the true correlation from the superpopulation. In all cases the use of H-S weights reduced the proportions to well below the expected .95 (compare Tables 7 and 3). Also, the extent of this bias became greater as the population standard deviation increased and as the number of studies in the meta-analysis and the sample sizes of those studies increased. This is not surprising because the H-S weights will be larger than the H-V weights, therefore, when they are replaced in equation (10) the resulting standard error of the mean z-transformed correlation will be smaller, and so too will be the resulting confidence intervals. This difference between the weights will increase as the between-studies variability estimate ($\hat{\tau}^2$)

increases, and as sample sizes increase. Given that the H-S weights reduce the accuracy of the H-V confidence intervals, the next question is whether the H-V weights improve the H-S confidence intervals. Table 8 shows these results. It is clear from comparing this table to Table 4, that using the H-V weights in the H-S confidence intervals makes virtually no difference: they show the same precision as before.

When the simulation was repeated but using a superpopulation based on *r* values, the findings were comparable to Simulation 2^{xi}: using the wrong weights made little difference to the average correlation from each method, however, using n_i in the H-V method dramatically reduced the proportion of confidence intervals that contained the true correlation (more so than when the superpopulation was based on *z*-transformed values), and using w_i^* in the H-S method had no noticeable effect on the proportion of confidence intervals containing the true correlation. As such the simulation process did not seem to interact with the weights in each method.

Discussion

Estimates of the True Correlation

This paper aimed to present extensive simulated data about the performance of the two most widely used random-effects methods of meta-analysis. Doing so has expanded earlier work (Field, 2001; Hall & Brannick, 2002; Schulze, 2004) to provide detailed information about when these methods can be trusted, and present comparative data about the relative strengths of the Hunter and Schmidt and Hedges methods. Some firm conclusions emerge from the initial predictions.

First, contrary to prediction 1 the estimates of the true correlation were not noticeably inaccurate when small numbers of studies were included in the meta-analysis. However, confidence intervals were: when small numbers of studies based on small sample sizes were included in the analysis confidence intervals were too wide

when the standard deviation of correlations was small ($\sigma_\rho \leq 0.08$); regardless of the sample size of studies, when the standard deviation of correlations was large ($\sigma_\rho \geq 0.16$) confidence intervals were too narrow when small numbers of studies were included in the meta-analysis.

Contrary to prediction 2, estimates of the true correlation from the H-S method were unaffected by the variability in correlations. Instead, estimates of the true correlation from the H-V method were affected by variability in correlations. Most noticeably, the overestimation of the true correlation by the H-V method was marked when the standard deviation of correlations was 0.16 or above. Also, the proportion of H-V confidence intervals containing the true correlation fell as variability between correlations increased. This was true also for the H-S method, but less so.

Contrary to prediction 3, the weights were not responsible for the differences between estimates of the true correlation: both methods appeared to produce similar estimates of the true correlation regardless of whether the correct weights were used or the weights from the opposite method. In terms of the confidence intervals from the two methods, weighting the correlation by the sample size in the H-V method did reduce the width of these intervals (and their accuracy). The reverse was not true: conducting the H-S method but using the weights from the H-V method had virtually no effect on the resulting confidence intervals.

In terms of prediction 4, H-S estimates seemed relatively unaffected by the standard deviation of correlations and the size of the true correlation in the superpopulation. In all simulation conditions H-S estimates of the true correlation were very accurate indeed and precision of the confidence intervals (although less precise than those from H-V's method) improved as correlation variability increased. However, H-V estimates of the true correlation were affected by the distribution of

correlations: estimates became less accurate and confidence intervals had lower coverage proportions as the standard deviation of correlations and the size of the true correlation became larger. These differences cannot be explained by the differing weights in the two methods (see above) and so these findings support Schulze's (2004) general position that the Fisher *z*-transformation is unnecessary (perhaps even unhelpful).

Finally, the way in which the superpopulation was simulated did have an effect on the results obtained: estimates of the true correlation using the H-V method were better when the superpopulation was based on *z*-transformed values than when it was based on *r*s. However, estimates from the H-S method were unaffected, and the profile of results remained unchanged by the simulation procedure: H-S estimates of the true correlation were as good as or better than those of the H-V method regardless of the simulation procedure. The data on the accuracy of confidence intervals was also unaffected in any substantive way by the simulation procedure: the H-V confidence intervals were more accurate when the superpopulation contained *z*-transformed scores, but were more accurate (generally—see below) than those from the H-S method even when the superpopulation was based on *r* values.

To sum up the findings about the accuracy of the estimates of the true effect, there was little to differentiate the H-V and H-S estimates of the true correlation when the standard deviation of correlations was small ($\sigma_\rho \leq 0.08$). However, H-V's method overestimated the true correlation when the true correlation was large ($\bar{\rho} \geq .3$) and the standard deviation of correlations was also large ($\sigma_\rho \geq 0.16$), and when the true correlation was small ($\bar{\rho} \geq .1$) and the standard deviation of correlations was at its maximum value ($\sigma_\rho = 0.32$). The H-S method produced very accurate estimates under all conditions (estimates were always within .011 of the true value).

The main deviation between the estimates from the two methods was when effect-size variability was large (σ_ρ = 0.32). Hall and Brannick (2002) believe that this level of variability is 'somewhat unrealistic' and so perhaps these findings are unimportant. However, Figure 1 shows that variability around this size *can* occur in real data, albeit relatively infrequently, and so these findings do have relevance—although it is worth remembering that this situation reflects the very extreme of what would occur in real-world data.

95% Confidence Intervals

In terms of coverage proportions of the 95% confidence intervals from the two methods, those from the H-V method were relatively more accurate than those from the H-S method. Coverage proportions for H-S's method were always too low and so these confidence intervals miss the true effect in the superpopulation more often than they should. Coverage proportions from the H-V method were generally on target, but deteriorated when a small number of studies (k = 5) were included in the meta-analysis and when the true effect, effect-size variability, and number of studies in the meta-analysis increased. This is likely to be a centering issue because these are exactly the combination of parameters that cause the average correlation from H-V to deviate from the true correlation (see, for example, Figure 1).

In some situations, confidence intervals based on the H-S method had better coverage: in general, as combinations of the number of studies included in the meta-analysis, the size of the true correlation and the variability of correlations increased, the confidence intervals from the H-S method were more likely to have the desired coverage than those from H-V. This suggests that the estimates of the standard error in the H-S method became more precise as the number of studies included in the

meta-analysis, the size of the true correlation and the variability of correlations, in combination, increased.

The standard error of the mean correlation from the H-S method underestimated the true standard error generally and this was especially true when the number of studies in the meta-analysis was small. This underestimation is because the variance of the observed *r*s (equation 13) is calculated using the sum of sample sizes in the denominator ($\sum_{i=1}^{k} n_i$), which is equivalent to the number of studies multiplied by the average sample size of all studies ($k\overline{N}$). In general, a less biased estimate of population variance is calculated using the number of observations minus 1 (for example, when estimating the population variance $n-1$ is used rather than n). In this case, the number of observations is the number of effect sizes, k, and so the denominator of equation 13 should be $(k-1)\overline{N}$ rather than $k\overline{N}$. The effect of this change would be to increase the estimated variance of observed *r*s, which in turn would increase the standard error of the mean correlation, which would widen the confidence intervals (see equation 17). This observation explains why the confidence intervals from H-S were too narrow when few studies were included (k is small).

The profile of results was unaffected by how the superpopulation was simulated, although the differences between methods were exaggerated when the superpopulation was based on *r*s. In addition, the weights used in the two methods did not appear to be responsible for the differences in the accuracy of estimates of the true correlation (although using sample size as the weight in H-V method certainly gave rise to poorer coverage proportions for the resulting confidence intervals).

Population Confidence Intervals

The population confidence intervals around the mean correlation (the boundaries within which 95% of average correlations fell) were smallest when the number of studies in the meta-analysis was large, and generally got bigger as the standard deviation of superpopulation correlations increased. Based on these population confidence intervals the following advice could be given: if a large average correlation is found ($\rho \geq .5$) then this will be within $\pm.1$ of the true value only if 20 or more studies were included in the meta-analysis (if $\sigma_\rho \leq 0.16$), or if 40 or more studies were included (if $\sigma_\rho > 0.16$). If a small or medium average correlation is found ($\rho = .1$ or .3) then this value will be within $\pm.1$ of the true value only when 40 (if $\sigma_\rho \leq 0.16$) or 80 (if $\sigma_\rho > 0.16$) or more studies are in the meta-analysis and the H-V method is used, but regardless of the correlation variability only 40 studies will be required if the H-S method is used. When there was no effect in the population and the standard deviation of population correlations was medium to extreme ($\sigma_\rho \geq 0.16$), the methods could erroneously detect a small average correlation unless more than about 40-80 studies (H-V) or 40 studies (H-S method) are included in the meta-analysis. Although the average sample size of the studies in the meta-analysis did make a difference (larger sample sizes produced tighter population confidence intervals), this difference was smaller than the effect of the number of studies in the meta-analysis and the standard deviation of correlations.

Comparisons with Previous Work

The results of the current study are consistent with those of Hall and Brannick (2002) and Field (2001) who found consistent overestimations of the average correlation from the H-V method. Hall and Brannick (2002) found overestimations ranging from 0 to .06 (and in general the overestimation increased as a function of

the standard deviation of correlations). Field (2001) found even greater overestimation: in the range of 0 to .20. In the current study, estimates from the H-V method produced a maximum overestimation of .031 (when the superpopulation was based on z_r values) and .052 (when the superpopulation was based on r). The latter value is based on a simulation procedure most similar to Field (2001) and Hall and Brannick (2002) and found a similar result.

Some cells of the design have approximately the same parameters examined in Field (2001) and can be directly compared. For example, when $\sigma_\rho = 0.16$, the number of studies in the meta-analysis was 30, and the superpopulation was based on r values, Field (2001) reported for superpopulation correlations of .1, .3, and .5 estimates from the H-V method of .144, .436, and .706 respectively. These are overestimations of .044, .136 and .206 respectively. In the present study, when $\sigma_\rho = 0.16$, the number of studies in the meta-analysis was 40, and the superpopulation was simulated using r values, the overestimations of population correlations of .1, .3, and .5, were (averaged across the different levels of sample size) .004, .013, and .025. Under the same conditions described above, Field (2001) reported for population correlations of .1, .3, and .5, estimates from the H-S method of .098, .293, and .478. These are deviations from the true value of -.002, -.007, and -.022 respectively. In the present study, the comparable estimates were -.001, -.003, and -.004. Thus, Field (2001) concluded that estimates of the average correlation from H-S were more accurate than from H-V — as did Hall and Brannick (2002). The current simulations tend to suggest the same conclusion; however, both methods produce more accurate average correlations than Field (2001) and, to a lesser extent, Hall and Brannick (2002) suggest.

The likely explanation for these small differences is the treatment of inadmissible values. In the current study, population correlations outside of the boundaries of -1 to

+1 were rejected (i.e., correlations were simply sampled until a value between -1 and +1 was obtained). In Field (2001) values outside of the upper and lower limit of 1 were capped; that is, replaced with a maximum value (0.999). Hall and Brannick (2002) used the same procedure as Field (2001) except values outside of ±.94 were capped at 0.94 (S. Hall, personal communication, 12th August, 2004). The result of a capping strategy is a build up of correlations at the very extreme of the distribution, which will be more pronounced as the superpopulation correlation and standard deviation increases. The effect of this build up of maximum values will be to drag the superpopulation mean upwards relative to when inadmissible correlations are rejected and regenerated. This will increase the H-V and H-S estimates of the true correlation. The overestimation in the H-V method is thus exaggerated. Also, because Field (2001) used a more extreme capping value than Hall and Brannick, there would be a greater build up of extreme values in his study, which explains the greater similarity between Hall and Brannick's results and those in the current study. In addition, the present study made attempts to compare the estimated mean correlations from the H-S and H-V methods to the actual mean correlation of the superpopulation (and not the hypothetical value set in the simulation). The build up of correlations when a capping strategy is used will change the mean correlation of those distributions, but unlike the current study, neither Field (2001) nor Hall and Brannick (2002) attempted to estimate the true mean of the superpopulation.

The current study also supports Schulze's broad conclusions (although not based on H-V's random-effects model for combining correlations) that the weights and z-transformation advocated by Hedges do not necessarily produce more accurate average effect sizes than those proposed by Hunter and Schmidt. Switching the weights in the two methods made very little difference to the average correlation.

Hall and Brannick's (2002) results differed from those of the current study in terms of the confidence intervals for the average effect size: they looked at *credibility intervals* from both methods and concluded that in most circumstances, the H-S method produced more accurate intervals. However, Hedges and Vevea (1998) do not advocate such intervals and instead provide equations for confidence intervals. The present study took the reverse approach and compared confidence intervals from both methods (which, conversely, Hunter and Schmidt do not advocate). The present study found a more complicated pattern of results than Hall and Brannick: the H-V confidence intervals were more accurate than those from the H-S method much of the time, but as combinations of the number of studies included in the meta-analysis, the size of the true correlation and the variability of correlations increased, the confidence intervals from the H-S method became more likely to be more accurate than those from H-V. However, even in these circumstances the H-S 95% confidence intervals (although an improvement on those from H-V) were still rather narrow containing only 94% of true correlation values.

Conclusions

Most researchers reading this article might expect an answer to the question in the title: Is the meta-analysis of correlation coefficients accurate when population correlations vary? Well, yes, by and large random-effects methods of meta-analysis produce accurate estimates of the true correlation. Although when the true correlation was large ($\bar{\rho} \geq .3$) and the standard deviation of correlations was also large ($\sigma_\rho \geq 0.16$), and when the true correlation was small ($\bar{\rho} \geq .1$) and the standard deviation of correlations was at its maximum value ($\sigma_\rho = 0.32$) the H-V method overestimated the true correlation, these overestimations were small (less than .052 above the true value). The H-S estimates were generally less biased than H-V estimates (less than

.011 below the true value). In terms of 95% confidence intervals, the H-S method only ever produced confidence intervals that contained the true correlation on 94% of occasions, and often much lower (as low as 83% when few studies were combined). Although 95% confidence intervals from H-V's method did, at times, contain 95% of true effect sizes, these confidence intervals could in certain circumstances be too wide (contained up to 96.2% of true correlations) or too narrow (contained 81.4% or 66% of true correlations depending on how the superpopulation was simulated). As such, researchers would need to make judgements about which method to use based on the size of the true correlation, the standard deviation of correlations (or estimates of these values), the number of studies being combined and the average sample size of studies in the meta-analysis.

References

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*, 1-26.

Becker, B. J. (1996). The generalizability of empirical research results. In C. P. Benbow and D. Lubinski (Eds.), *Intellectual talent: Psychological and social issues*, Baltimore: Johns Hopkins University Press, 363-383.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd Ed.). New York: Academic Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods*, *6*, 161–18.

Field, A. P. (2003a). The problems in using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics*, *2*, 77-96.

Field, A. P. (2003b). Can meta-analysis be trusted? *The Psychologist*, *16*, 642-645.

Field, A. P. (2005). *Discovering statistics using SPSS (second edition)*. London: Sage.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 3–32.

Friedman, L. (2000). Estimators of random effects variance components in meta-analysis. *Journal of Educational and Behavioural Statistics*, *25,* 1-12.

GAUSS for Windows 4.0 [Computer software]. (1984–2002). Maple Valley, WA: Aptech Systems, Inc.

Glass, G. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3–8.

Hall, S. M., & Brannick, M. T. (2002). Comparison of two random-effects methods of meta-analysis. *Journal of Applied Psychology*, *87*, 377–389.

Hedges, L. V. (1992). Meta-Analysis. *Journal of Educational Statistics*, *17*, 279-296.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V. & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, *6*, 203-217.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486-504.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (1990b). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology*, *75*, 334–349.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative knowledge in psychology. *International Journal of Selection and Assessment*, *8*, 275–292.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings (Second edition)*. Newbury Park, CA: Sage.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.

Johnson, B. T., Mullen, B., & Salas, E. (1995). Comparison of three major meta-analytic approaches. *Journal of Applied Psychology*, *80*, 94-106.

Mooney, C. Z. (1997). *Monte Carlo Simulation* (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-116). Thousand Oaks, CA: Sage.

National Research Council (1992). *Combining information: Statistical issues and opportunities for research*. Washington, D.C.: National Academy Press.

Osburn, H. G., & Callender, J. (1992). A note on the sampling variance of the mean uncorrected correlation in meta-analysis and validity generalization. *Journal of Applied Psychology*, *77*, 115-122.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods*, *3*, 354-379.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (revised). Newbury Park, CA: Sage.

Rosenthal, R., & Rubin, D. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavior and Brain Sciences*, *3*, 377-415.

Sackett, P. R., Harris, M. M., & Orr, J. M. (1986). On seeking moderator variables in the meta-analysis of correlational data: A Monte Carlo investigation of statistical power and resistance to Type I error. *Journal of Applied Psychology*, *71*, 302–310.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*, 529-540.

Schmidt, F. L., & Hunter, J. E. (1999). Comparison of three meta-analysis methods revisited: An analysis of Johnson, Mullen, and Salas (1995). *Journal of Applied Psychology*, *84* (1), 144-148.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology*, *61*, 473-485.

Schulze, R. (2004). *Meta-analysis: a comparison of approaches*. Cambridge, MA: Hogrefe & Huber.

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's Z transformation be used? *Journal of Applied Psychology*, *72*, 146-148.

Strube, M. J. (1988). Averaging correlation coefficients: Influence of heterogeneity and set size. *Journal of Applied Psychology*, *73*, 559-568.

Takkouche, B., Cadarso-Suarez, C., & Spiegelman, D. (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology*, *150*, 206-215.

Wolf, F. M. (1986). *Meta-analysis*. Sage university paper series on quantitative applications in the social sciences, 07–061. Newbury Park, CA: Sage.

Author note

Correspondence concerning this article should be addressed to Andy P. Field, Department of Psychology, University of Sussex, Falmer, Brighton, East Sussex, BN1 9QH. Electronic mail may be sent to andyf@sussex.ac.uk.

I am grateful to Steven Hall and Michael Brannick for their co-operation in discussing their simulation methods and to Huy Le for doing an independent check of my simulation code. Seven reviewers provided comments that led to vast improvements in this paper. In particular, Frank Schmidt's reviews were constructive, detailed and educational; he has been extremely generous with his time, ideas and technical and theoretical advice. This paper (and my understanding of meta-analysis) have improved, $r = 1$, because of his altruism.

Table 1: Actual means and SDs of the distributions of ρ simulated by an *Fz* Transformation

Simulated $\overline{\rho}$	SD(Fz)			
	0.04	0.08	0.16	0.32
	Actual Mean of Distribution			
.0	.0000	.0000	.0000	.0002
.1	.0995	.09900	.0972	.0912
.3	.2909	.2900	.2848	.2679
.5	.4615	.4598	.4531	
.8	.6634			
	Actual SD of Distribution			
.0	0.0399	0.0795	0.1561	0.2932
.1	0.0395	0.0787	0.1547	0.2910
.3	0.0366	0.0729	0.1440	0.2749
.5	0.0315	0.0629	0.1253	
.8	0.0224			

Table 2: Values of skew and kurtosis in population distributions (* indicates that the associated *z*-test is significant at p < .001, SE_{Skew} = .011, $SE_{Kurtosis}$ = .022 throughout)

		Superpopulation			
		Based on *r*		Based on *z*	
σ_ρ	$\overline{\rho}$	Skew	Kurtosis	Skew	Kurtosis
0.04	.0	0.012	-0.009	-0.010	0.012
	.1	0.000	-0.025	-0.032	-0.016
	.3	0.008	-0.062	-0.067*	-0.008
	.5	-0.002	0.024	-0.110*	0.023
	.8	0.004	-0.017	-0.167*	0.021
0.08	.0	-0.004	-0.043	-0.002	-0.047
	.1	0.013	0.047	-0.029	-0.070*
	.3	0.013	0.033	-0.141*	0.032
	.5	-0.002	0.009	-0.214*	0.027
0.16	.0	-0.021	0.003	0.005	-0.184*
	.1	0.010	0.002	-0.096*	-0.159*
	.3	-0.007	0.010	-0.247*	-0.068
	.5	-0.024	-0.076*	-0.412*	0.127*
0.32	.0	-0.001	-0.136*	-0.005	-0.512*
	.1	-0.051*	-0.142*	-0.133*	-0.468*
	.3	-0.165*	-0.260*	-0.388*	-0.292*

Table 3: Proportion of confidence intervals based on the H-V method that actually contained the true correlation when the superpopulation is based on z_r values (True Effect = the mean correlation in the superpopulation)

		Average Sample Size																			
		True effect = 0				True effect = .1				True effect = .3				True effect = .5				True effect = .66			
Number of Studies		20	40	80	160	20	40	80	160	20	40	80	160	20	40	80	160	20	40	80	160
		$\sigma_p = 0.04$																			
k	5	.962	.958	.955	.945	.962	.959	.954	.946	.962	.958	.954	.946	.960	.959	.955	.944	.958	.957	.954	.945
	10	.959	.957	.954	.945	.959	.958	.954	.945	.959	.956	.954	.947	.956	.957	.952	.947	.954	.955	.951	.945
	20	.958	.956	.952	.947	.957	.956	.952	.946	.955	.954	.953	.947	.952	.953	.950	.945	.946	.949	.949	.944
	40	.957	.953	.950	.947	.956	.955	.950	.947	.952	.951	.949	.946	.943	.948	.947	.945	.928	.939	.943	.940
	80	.954	.953	.950	.947	.954	.953	.949	.947	.944	.946	.946	.945	.927	.938	.942	.942	.896	.922	.932	.937
	160	.953	.951	.950	.947	.951	.949	.948	.947	.930	.939	.943	.944	.896	.922	.933	.937	.833	.889	.912	.924
		$\sigma_p = 0.08$																			
	5	.955	.947	.933	.914	.955	.946	.932	.914	.954	.947	.933	.914	.953	.944	.932	.914				
	10	.955	.946	.936	.926	.954	.947	.936	.925	.952	.946	.936	.926	.950	.944	.935	.926				
	20	.954	.948	.942	.936	.952	.947	.940	.936	.949	.944	.938	.934	.944	.942	.936	.934				
	40	.951	.948	.943	.944	.950	.947	.943	.943	.945	.943	.940	.941	.934	.937	.934	.936				
	80	.950	.947	.946	.945	.949	.946	.946	.945	.936	.938	.940	.942	.916	.924	.929	.933				
	160	.952	.948	.947	.948	.946	.946	.946	.947	.922	.931	.935	.938	.876	.902	.915	.921				
		$\sigma_p = 0.16$																			
	5	.936	.915	.895	.883	.935	.915	.896	.884	.934	.913	.895	.885	.931	.912	.894	.881				
	10	.938	.926	.919	.917	.937	.927	.919	.918	.936	.923	.917	.918	.929	.919	.912	.911				
	20	.940	.935	.935	.934	.939	.935	.934	.934	.933	.932	.930	.930	.925	.922	.922	.923				
	40	.943	.942	.943	.943	.942	.941	.941	.942	.931	.932	.933	.933	.910	.916	.918	.918				
	80	.946	.947	.945	.947	.944	.944	.945	.945	.920	.925	.926	.928	.879	.890	.896	.899				
	160	.948	.948	.948	.949	.942	.943	.942	.945	.897	.907	.910	.911	.814	.837	.845	.851				
		$\sigma_p = 0.32$																			
	5	.897	.886	.879	.879	.896	.883	.880	.875	.894	.878	.874	.873								
	10	.921	.918	.919	.918	.918	.916	.917	.918	.910	.908	.907	.909								
	20	.936	.935	.935	.938	.933	.934	.933	.936	.917	.920	.919	.919								
	40	.942	.942	.942	.943	.938	.939	.939	.941	.910	.912	.913	.914								
	80	.946	.947	.947	.947	.940	.939	.940	.940	.881	.884	.887	.885								
	160	.949	.948	.948	.949	.934	.934	.935	.935	.822	.825	.825	.827								

Table 4: Proportion of confidence intervals based on the H-S method that actually contained the true correlation when the superpopulation is based on z_r values (True Effect = the mean correlation in the superpopulation)

		Average Sample Size																			
		True effect = 0				True effect = .1				True effect = .3				True effect = .5				True effect = .66			
Number of Studies		20	40	80	160	20	40	80	160	20	40	80	160	20	40	80	160	20	40	80	160
		σ_ρ = 0.04																			
	5	.843	.846	.844	.843	.845	.846	.844	.844	.845	.845	.845	.846	.843	.843	.845	.840	.842	.843	.845	.841
	10	.903	.903	.903	.900	.904	.905	.903	.901	.904	.904	.904	.902	.903	.904	.903	.903	.905	.903	.905	.901
	20	.929	.928	.927	.927	.928	.929	.927	.928	.928	.928	.928	.929	.928	.928	.927	.927	.929	.928	.927	.928
	40	.941	.938	.938	.938	.941	.941	.938	.939	.938	.938	.938	.939	.935	.937	.939	.938	.932	.934	.937	.937
	80	.945	.945	.944	.943	.945	.946	.943	.943	.939	.941	.942	.943	.929	.938	.941	.941	.912	.927	.936	.940
	160	.947	.946	.947	.945	.947	.945	.946	.945	.932	.939	.942	.944	.908	.928	.938	.942	.861	.906	.928	.939
		σ_ρ = 0.08																			
	5	.847	.844	.842	.841	.844	.842	.842	.840	.842	.845	.841	.841	.842	.841	.842	.841				
	10	.904	.903	.900	.899	.903	.903	.900	.899	.902	.902	.902	.901	.901	.902	.902	.900				
	20	.929	.928	.928	.925	.928	.928	.926	.925	.927	.926	.925	.924	.926	.928	.926	.927				
	40	.939	.940	.936	.937	.938	.939	.937	.937	.938	.938	.937	.936	.936	.938	.936	.936				
	80	.944	.943	.942	.940	.944	.943	.942	.941	.940	.941	.942	.942	.931	.937	.940	.942				
	160	.949	.946	.944	.944	.945	.946	.944	.944	.933	.940	.943	.943	.911	.930	.938	.942				
k		σ_ρ = 0.16																			
	5	.841	.840	.837	.835	.840	.840	.838	.836	.840	.839	.838	.838	.838	.837	.836	.834				
	10	.902	.899	.896	.897	.901	.899	.897	.898	.901	.899	.897	.898	.900	.897	.895	.895				
	20	.926	.925	.923	.922	.925	.925	.923	.921	.925	.926	.923	.923	.927	.924	.922	.921				
	40	.936	.936	.934	.934	.937	.935	.934	.934	.937	.936	.935	.934	.935	.936	.935	.934				
	80	.942	.942	.939	.939	.942	.942	.940	.940	.939	.941	.938	.940	.934	.939	.939	.940				
	160	.946	.944	.942	.943	.944	.943	.941	.942	.937	.941	.941	.942	.921	.937	.940	.942				
		σ_ρ = 0.32																			
	5	.836	.836	.835	.835	.835	.834	.836	.831	.834	.830	.830	.831								
	10	.898	.895	.897	.895	.896	.895	.895	.896	.895	.894	.891	.891								
	20	.924	.922	.921	.922	.923	.923	.922	.924	.922	.921	.920	.919								
	40	.934	.934	.932	.933	.934	.934	.934	.933	.935	.935	.933	.932								
	80	.940	.939	.939	.938	.940	.939	.939	.939	.939	.939	.937	.938								
	160	.943	.942	.939	.941	.941	.941	.940	.941	.941	.942	.940	.940								

Table 5: Proportion of confidence intervals based on the H-V method that actually contained the true correlation when the superpopulation is based on *r* values

		Average Sample Size																			
		True effect = 0				True effect = .1				True effect = .3				True effect = .5				True effect = .8			
Number of Studies		20	40	80	160	20	40	80	160	20	40	80	160	20	40	80	160	20	40	80	160
		$\sigma_\rho = 0.04$																			
k	5	.962	.958	.954	.947	.962	.959	.953	.945	.961	.958	.952	.943	.959	.954	.948	.934	.941	.930	.911	.894
	10	.960	.957	.953	.946	.960	.958	.953	.946	.958	.956	.952	.944	.954	.952	.947	.937	.932	.925	.919	.917
	20	.959	.955	.952	.947	.957	.956	.952	.946	.956	.954	.949	.944	.950	.949	.944	.941	.913	.917	.917	.921
	40	.957	.955	.951	.948	.956	.953	.950	.947	.950	.951	.948	.944	.938	.943	.942	.939	.875	.891	.904	.911
	80	.954	.953	.950	.947	.954	.951	.950	.947	.943	.945	.945	.944	.918	.931	.935	.938	.801	.843	.862	.874
	160	.954	.952	.949	.947	.950	.949	.948	.947	.929	.939	.941	.942	.882	.909	.923	.930	.649	.732	.770	.790
		$\sigma_\rho = 0.08$																			
	5	.955	.946	.933	.914	.955	.947	.930	.914	.953	.941	.928	.910	.947	.934	.916	.898				
	10	.954	.947	.937	.926	.954	.947	.936	.926	.951	.943	.933	.924	.943	.935	.925	.921				
	20	.953	.948	.939	.937	.953	.945	.941	.935	.947	.942	.937	.936	.934	.930	.929	.930				
	40	.952	.946	.943	.942	.950	.946	.943	.942	.943	.940	.939	.940	.922	.927	.928	.932				
	80	.952	.947	.946	.947	.948	.947	.944	.946	.933	.936	.939	.941	.894	.909	.916	.920				
	160	.950	.947	.948	.947	.947	.946	.947	.947	.916	.926	.931	.935	.840	.871	.885	.894				
		$\sigma_\rho = 0.16$																			
	5	.934	.914	.897	.884	.933	.912	.894	.884	.928	.910	.893	.885	.915	.898	.888	.883				
	10	.937	.926	.922	.919	.938	.926	.920	.920	.930	.923	.919	.917	.917	.914	.914	.916				
	20	.941	.936	.936	.935	.940	.934	.935	.936	.930	.929	.931	.931	.910	.914	.919	.920				
	40	.943	.942	.943	.943	.942	.941	.942	.943	.926	.929	.931	.932	.880	.889	.896	.900				
	80	.947	.947	.946	.947	.942	.944	.945	.945	.911	.918	.919	.922	.807	.825	.831	.837				
	160	.948	.948	.948	.949	.941	.943	.943	.944	.876	.888	.888	.895	.660	.686	.694	.700				
		$\sigma_\rho = 0.32$																			
	5	.897	.890	.885	.884	.899	.888	.883	.883	.895	.887	.886	.885								
	10	.921	.920	.919	.922	.919	.918	.921	.921	.914	.915	.917	.917								
	20	.935	.936	.936	.936	.933	.934	.935	.934	.916	.918	.920	.920								
	40	.943	.944	.945	.944	.938	.938	.939	.939	.892	.893	.896	.897								
	80	.948	.946	.947	.947	.933	.934	.934	.936	.823	.826	.829	.831								
	160	.948	.949	.948	.948	.922	.920	.922	.922	.680	.685	.688	.689								

Table 6: Proportion of confidence intervals based on the H-S method that actually contained the true correlation when the superpopulation is based on *r* values

		Average Sample size																			
		True effect = 0				True effect = .1				True effect = .3				True effect = .5				True effect = .8			
Number of studies		20	40	80	160	20	40	80	160	20	40	80	160	20	40	80	160	20	40	80	160
		$\sigma_\rho = 0.04$																			
k	5	.846	.844	.843	.846	.847	.846	.843	.843	.845	.843	.843	.843	.843	.844	.843	.841	.838	.841	.838	.837
	10	.903	.904	.903	.903	.905	.903	.901	.901	.904	.902	.904	.901	.902	.904	.902	.902	.902	.900	.898	.899
	20	.929	.927	.928	.927	.928	.929	.928	.927	.928	.929	.928	.926	.929	.929	.928	.928	.926	.926	.924	.925
	40	.940	.940	.939	.940	.941	.939	.938	.939	.938	.939	.939	.938	.935	.937	.938	.937	.929	.933	.935	.935
	80	.945	.945	.944	.943	.945	.944	.944	.944	.939	.941	.942	.943	.926	.936	.940	.942	.908	.928	.937	.939
	160	.948	.948	.946	.945	.946	.945	.946	.945	.930	.940	.943	.944	.902	.926	.936	.943	.848	.907	.929	.938
		$\sigma_\rho = 0.08$																			
	5	.846	.843	.842	.840	.841	.838	.838	.836	.844	.842	.843	.842	.839	.842	.838	.837				
	10	.904	.903	.902	.900	.901	.901	.898	.899	.903	.901	.901	.900	.902	.901	.900	.900				
	20	.929	.929	.925	.926	.926	.924	.923	.924	.928	.926	.926	.927	.926	.925	.924	.924				
	40	.939	.937	.937	.936	.937	.937	.934	.935	.938	.936	.937	.935	.935	.937	.936	.936				
	80	.946	.943	.942	.942	.942	.942	.940	.939	.937	.941	.942	.941	.929	.936	.939	.940				
	160	.947	.945	.945	.942	.944	.943	.941	.942	.932	.939	.941	.943	.906	.930	.937	.942				
		$\sigma_\rho = 0.16$																			
	5	.844	.840	.839	.837	.841	.838	.838	.836	.840	.840	.838	.838	.837	.836	.837	.836				
	10	.900	.900	.900	.898	.901	.901	.898	.899	.899	.899	.899	.897	.899	.898	.896	.896				
	20	.927	.924	.924	.921	.926	.924	.923	.924	.925	.924	.923	.921	.925	.922	.923	.920				
	40	.936	.936	.935	.935	.937	.937	.934	.935	.936	.935	.935	.934	.935	.935	.934	.934				
	80	.943	.941	.940	.940	.942	.942	.940	.939	.939	.939	.939	.938	.935	.938	.939	.939				
	160	.945	.944	.942	.943	.944	.943	.941	.942	.937	.941	.940	.941	.924	.936	.939	.941				
		$\sigma_\rho = 0.32$																			
	5	.835	.838	.836	.836	.838	.836	.834	.835	.836	.835	.836	.835								
	10	.897	.897	.895	.897	.896	.895	.896	.895	.895	.895	.895	.895								
	20	.922	.922	.921	.920	.923	.921	.921	.920	.922	.921	.922	.920								
	40	.934	.933	.934	.933	.935	.932	.933	.933	.935	.932	.935	.932								
	80	.942	.938	.938	.938	.939	.939	.938	.939	.940	.939	.938	.935								
	160	.941	.941	.940	.941	.944	.941	.941	.941	.939	.942	.942	.939								

Table 7: Proportion of confidence intervals based on the H-V method, but with H-S weights, that contained the true correlation when the superpopulation is based on z_r values (True Effect = the mean correlation in the superpopulation)

		Average Sample Size																			
		True effect = 0				True effect = .1				True effect = .3				True effect = .5				True effect = .66			
Number of Studies		20	40	80	160	20	40	80	160	20	40	80	160	20	40	80	160	20	40	80	160
		$\sigma_\rho = 0.04$																			
	5	.924	.931	.930	.915	.925	.932	.929	.916	.923	.932	.930	.916	.922	.932	.929	.914	.921	.930	.930	.914
	10	.923	.933	.930	.914	.924	.933	.929	.915	.924	.930	.929	.917	.919	.931	.928	.915	.917	.929	.927	.914
	20	.925	.932	.929	.915	.925	.932	.928	.915	.920	.930	.930	.916	.917	.928	.926	.914	.908	.924	.924	.912
	40	.926	.931	.928	.915	.924	.933	.928	.914	.918	.929	.926	.914	.907	.925	.924	.913	.886	.913	.920	.907
	80	.925	.932	.929	.915	.923	.933	.928	.915	.911	.925	.925	.913	.887	.915	.919	.908	.848	.896	.908	.901
	160	.925	.931	.930	.914	.921	.929	.928	.913	.894	.918	.921	.910	.849	.897	.910	.900	.771	.857	.885	.884
		$\sigma_\rho = 0.08$																			
	5	.911	.909	.881	.823	.911	.906	.881	.823	.910	.907	.879	.825	.908	.905	.878	.825				
	10	.913	.907	.880	.823	.912	.908	.879	.822	.908	.907	.880	.824	.905	.904	.878	.821				
	20	.913	.908	.882	.823	.911	.908	.880	.822	.906	.903	.877	.820	.901	.900	.874	.818				
	40	.912	.909	.880	.824	.909	.907	.880	.822	.902	.901	.876	.819	.886	.893	.867	.811				
	80	.912	.906	.881	.823	.910	.906	.879	.822	.892	.894	.871	.814	.864	.875	.855	.801				
	160	.913	.906	.878	.823	.907	.904	.877	.821	.874	.883	.861	.803	.815	.845	.828	.773				
k		$\sigma_\rho = 0.16$																			
	5	.863	.816	.723	.604	.861	.816	.725	.604	.861	.814	.723	.602	.857	.812	.725	.600				
	10	.863	.814	.724	.605	.860	.815	.723	.604	.857	.810	.720	.599	.850	.802	.713	.592				
	20	.862	.815	.724	.603	.859	.815	.721	.603	.850	.806	.714	.594	.837	.791	.701	.583				
	40	.861	.813	.726	.600	.858	.812	.722	.599	.841	.796	.706	.584	.810	.767	.676	.557				
	80	.862	.815	.725	.601	.856	.810	.718	.599	.822	.777	.687	.565	.759	.722	.634	.518				
	160	.861	.815	.724	.602	.851	.805	.713	.592	.781	.741	.653	.533	.663	.637	.552	.445				
		$\sigma_\rho = 0.32$																			
	5	.708	.598	.473	.353	.708	.598	.471	.354	.703	.592	.464	.349								
	10	.708	.596	.473	.354	.708	.596	.468	.353	.691	.581	.456	.340								
	20	.708	.599	.468	.355	.707	.597	.464	.350	.678	.568	.447	.336								
	40	.705	.597	.467	.353	.699	.591	.464	.346	.652	.544	.422	.315								
	80	.709	.595	.467	.352	.695	.584	.456	.343	.593	.493	.378	.284								
	160	.710	.595	.466	.353	.681	.570	.446	.333	.500	.403	.305	.226								

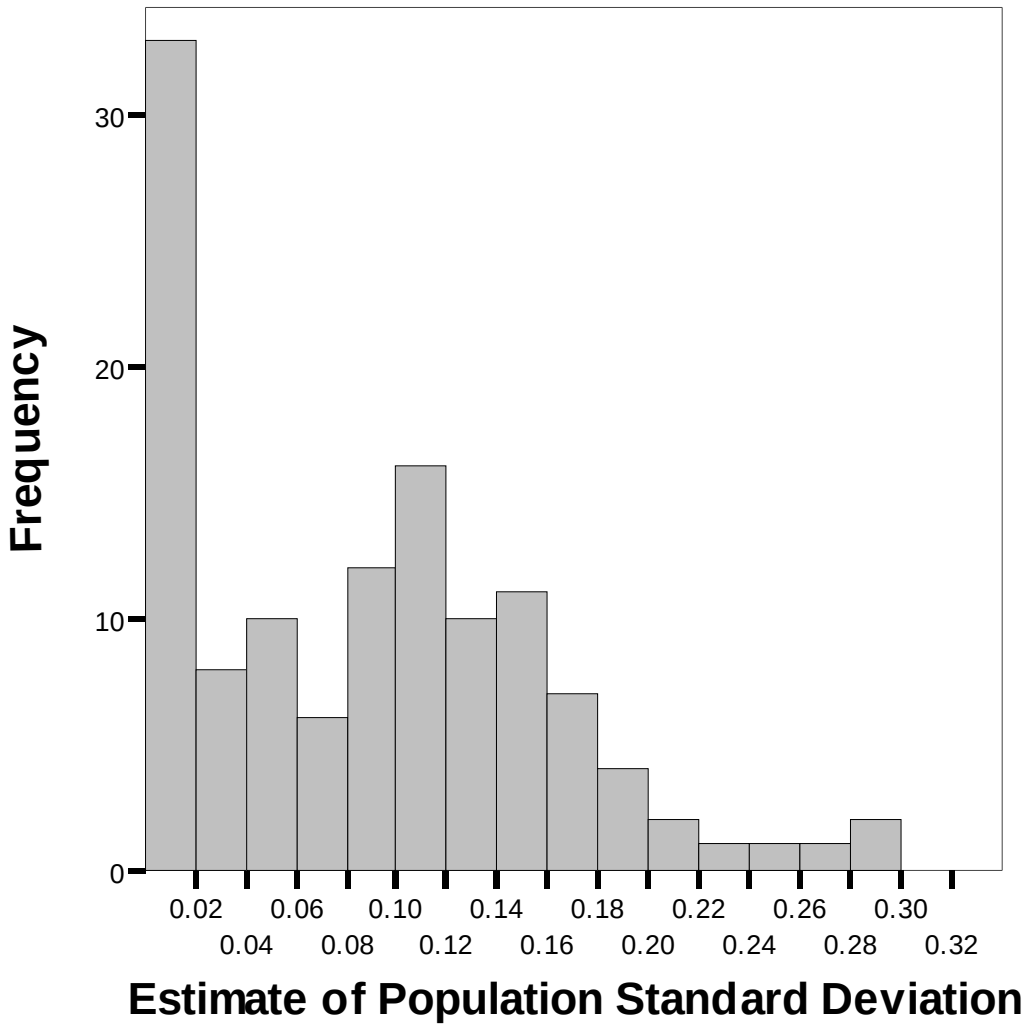
Table 8: Proportion of confidence intervals based on the H-S method, but with H-V weights, that contained the true correlation when the superpopulation is based on z_r values

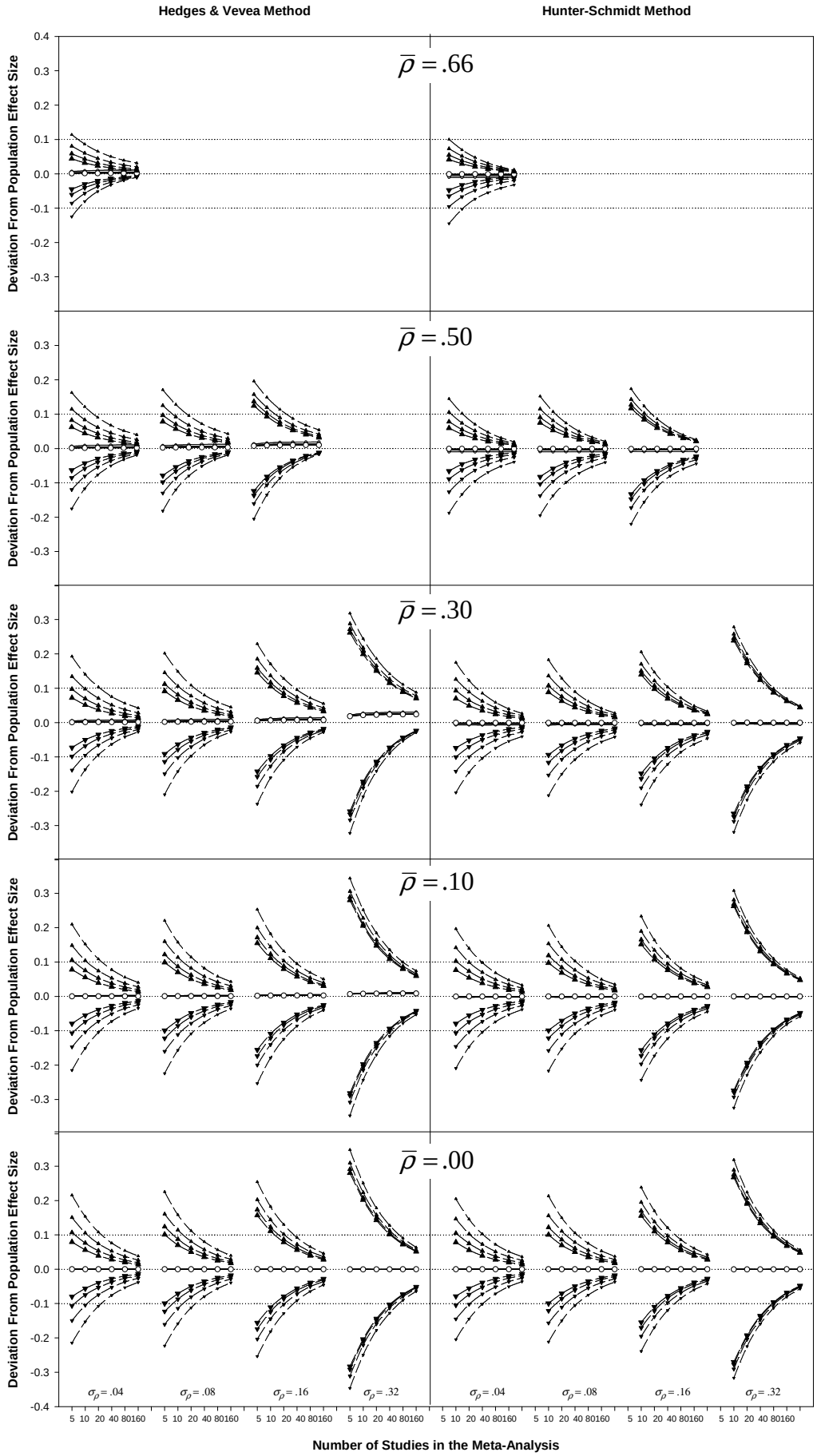
		Average Sample Size																			
		True effect = 0				True effect = .1				True effect = .3				True effect = .5				True effect = .66			
Number of studies		20	40	80	160	20	40	80	160	20	40	80	160	20	40	80	160	20	40	80	160
		$\sigma_\rho = 0.04$																			
k	5	.841	.845	.844	.843	.843	.845	.844	.845	.843	.844	.845	.846	.841	.842	.845	.841	.840	.843	.845	.841
	10	.901	.902	.903	.900	.902	.904	.904	.902	.902	.904	.904	.903	.902	.904	.904	.904	.903	.902	.905	.902
	20	.928	.927	.928	.928	.926	.928	.927	.928	.927	.927	.929	.929	.927	.927	.927	.927	.927	.927	.928	.929
	40	.939	.938	.939	.939	.939	.941	.938	.940	.936	.938	.938	.939	.933	.937	.939	.939	.930	.933	.937	.938
	80	.943	.945	.945	.944	.943	.946	.943	.945	.938	.941	.943	.943	.927	.938	.941	.942	.911	.926	.937	.941
	160	.946	.946	.948	.946	.945	.945	.946	.946	.930	.939	.943	.945	.907	.928	.938	.943	.861	.906	.928	.940
		$\sigma_\rho = 0.08$																			
	5	.844	.844	.843	.843	.841	.841	.842	.842	.839	.844	.842	.844	.839	.841	.843	.844				
	10	.902	.902	.901	.902	.901	.903	.901	.901	.901	.902	.903	.903	.899	.902	.903	.902				
	20	.928	.928	.930	.928	.926	.928	.928	.928	.925	.926	.926	.927	.925	.928	.927	.930				
	40	.938	.940	.938	.940	.937	.939	.939	.940	.936	.938	.939	.940	.935	.938	.938	.939				
	80	.943	.943	.944	.943	.943	.943	.944	.944	.938	.941	.944	.945	.929	.937	.942	.945				
	160	.948	.946	.946	.947	.944	.946	.946	.947	.932	.941	.946	.946	.909	.930	.940	.945				
		$\sigma_\rho = 0.16$																			
	5	.840	.841	.841	.842	.839	.841	.841	.842	.839	.840	.842	.844	.837	.838	.840	.840				
	10	.901	.902	.901	.903	.901	.901	.901	.903	.901	.900	.902	.904	.899	.899	.899	.900				
	20	.926	.927	.928	.928	.925	.927	.928	.928	.924	.928	.927	.928	.927	.926	.927	.928				
	40	.937	.939	.939	.940	.937	.939	.939	.940	.937	.939	.939	.940	.935	.939	.940	.939				
	80	.943	.946	.944	.945	.943	.944	.945	.945	.940	.944	.943	.946	.935	.942	.944	.945				
	160	.946	.947	.947	.948	.945	.947	.946	.948	.938	.944	.946	.948	.922	.939	.944	.948				
		$\sigma_\rho = 0.32$																			
	5	.837	.841	.843	.844	.837	.840	.843	.840	.836	.835	.838	.839								
	10	.901	.902	.904	.903	.899	.901	.902	.905	.898	.899	.899	.900								
	20	.927	.928	.928	.931	.927	.929	.929	.931	.925	.927	.927	.927								
	40	.938	.940	.939	.940	.939	.940	.940	.941	.939	.940	.939	.939								
	80	.944	.945	.945	.945	.945	.945	.945	.945	.943	.945	.944	.945								
	160	.948	.947	.947	.948	.945	.947	.948	.947	.946	.948	.946	.947								

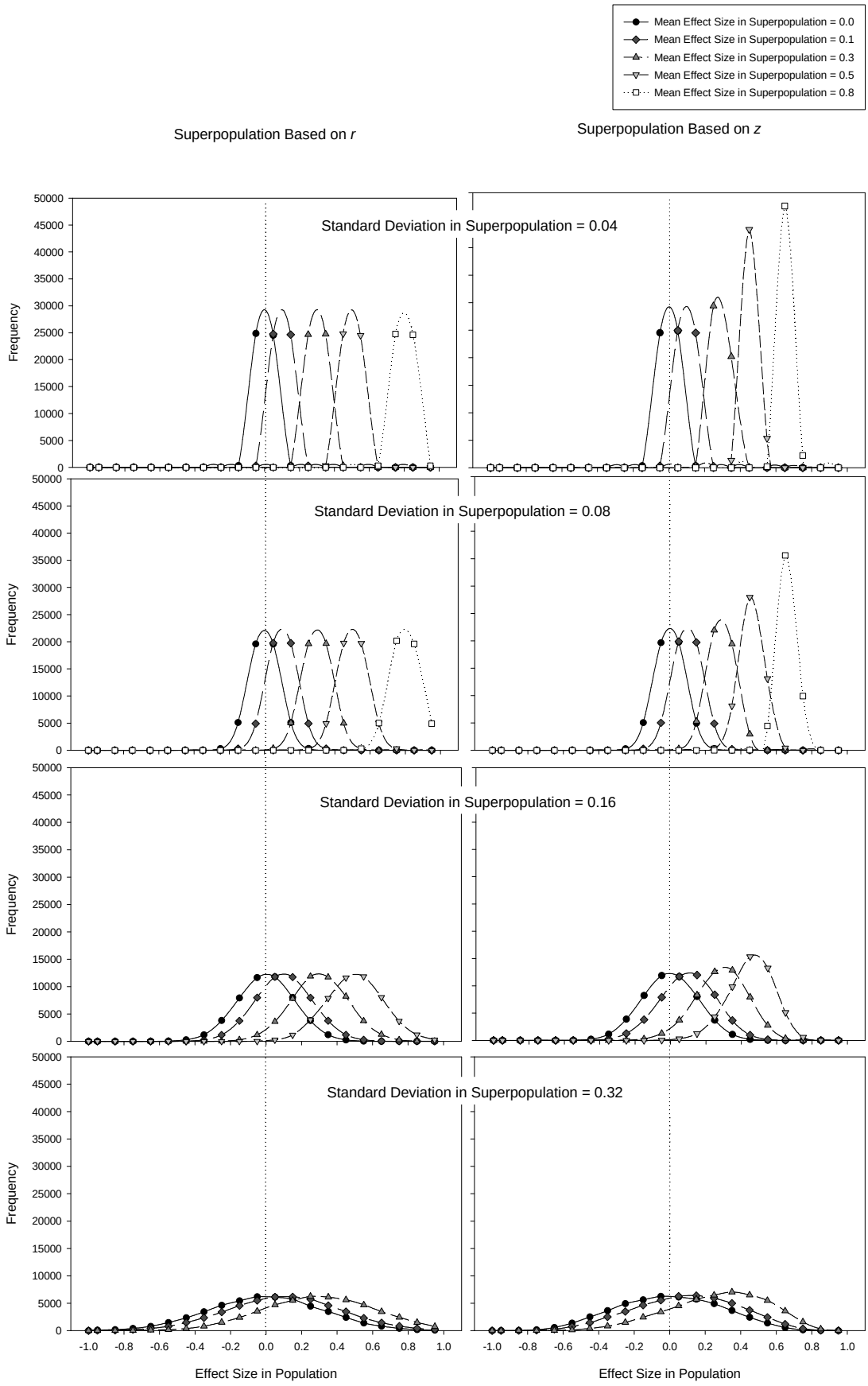
FIGURES

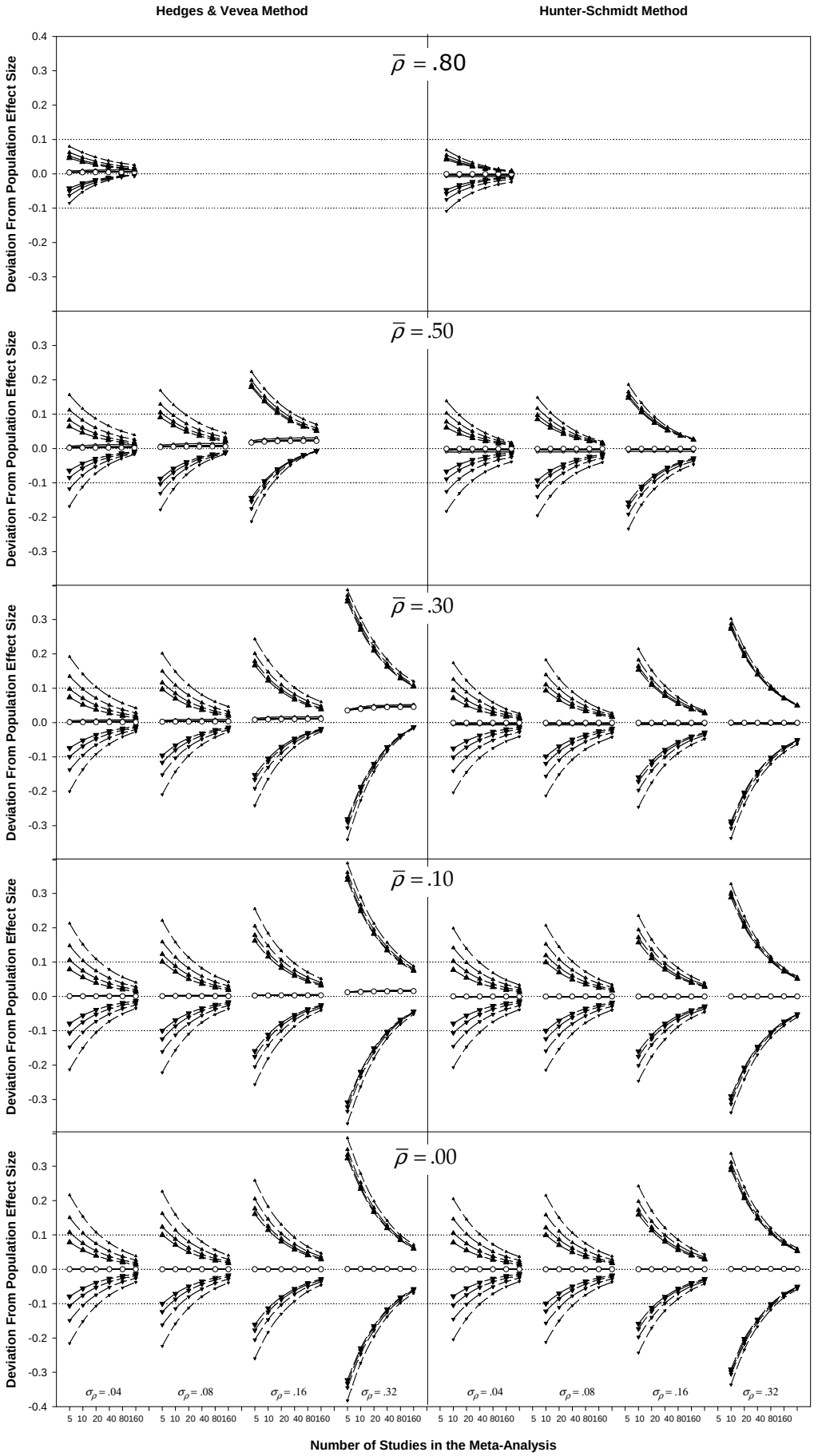
- Figure 1: Histogram showing the frequency of different levels of variability between effect sizes in meta-analytic studies using correlation coefficients published in Psychological Bulletin 1997-2002. The estimate of the population standard deviation was calculated using Hunter and Schmidt's method (the square root of equation (15)).
- Figure 2: The deviation from the true value of average correlations (circles) and the lower and upper boundaries of the 95% confidence interval (triangles) from Hedges-Vevea and Hunter-Schmidt methods of meta-analysis when the superpopulation is based on *z*-transformed values. The average sample sizes of studies in the meta-analysis are shown by the size of circles and triangles (smaller circles and triangles represent smaller average sample sizes). Average correlations and confidence intervals were compared to the values in Table 1.
- Figure 3: Frequency of population correlations as the average correlation in the superpopulation and its standard deviation varies. Graphs are shown when z-transformed values of *r* were used to model the superpopulation, and when the superpopulation was based on values of *r* but with inadmissible values replaced.
- Figure 4: The deviation from the true value of average correlations (circles) and the lower and upper boundaries of the 95% confidence interval (triangles) from Hedges-Vevea and Hunter-Schmidt methods of meta-analysis when the superpopulation is based on *r* values. The average sample sizes of studies in the meta-analysis are shown by the size of circles and triangles (smaller circles and triangles represent smaller average sample sizes).

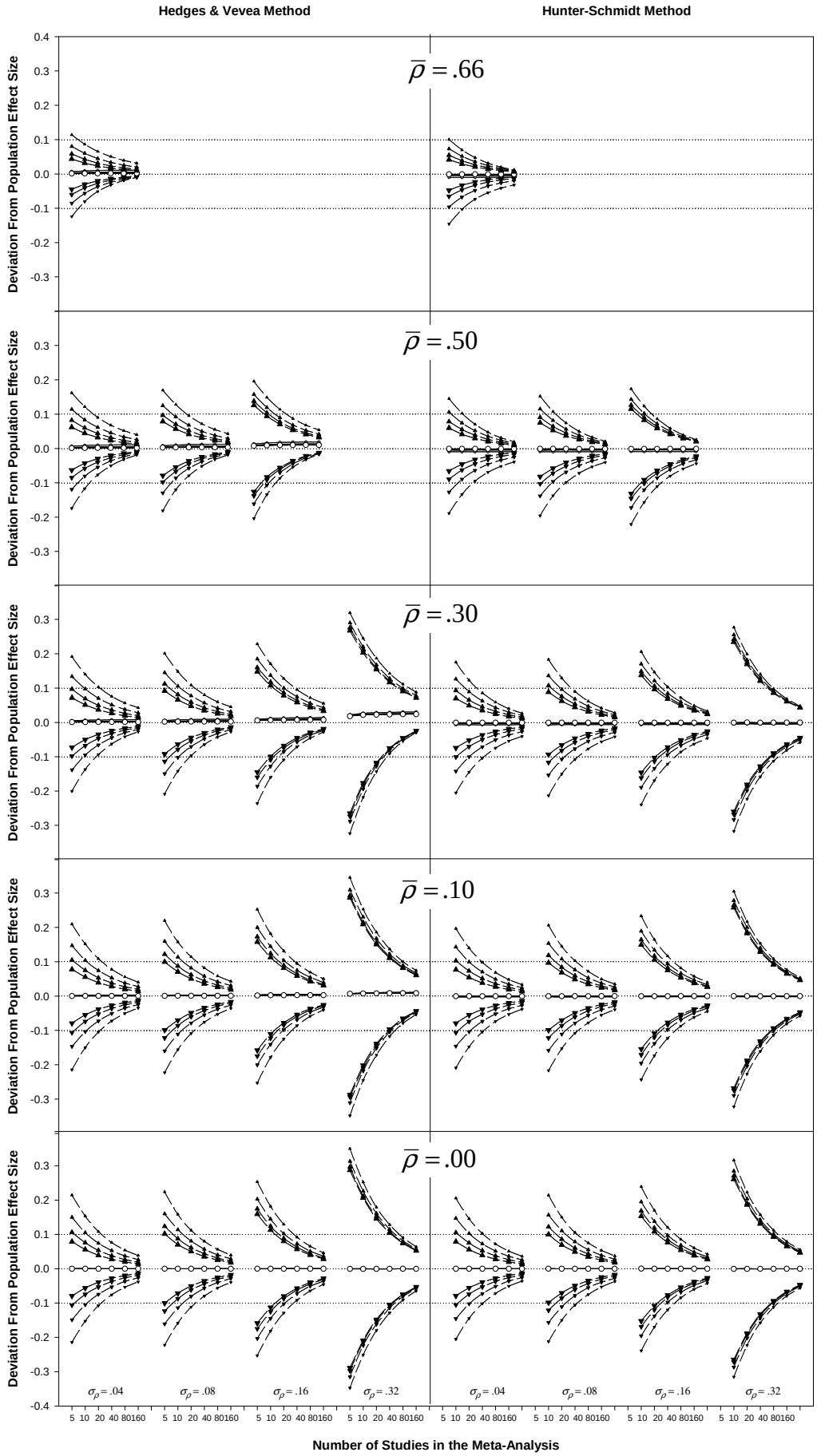
- Figure 5: The deviation from the true value of average correlations (circles) and the lower and upper boundaries of the 95% confidence interval (triangles) from Hedges-Vevea and Hunter-Schmidt methods of meta-analysis when the wrong weights are used (the superpopulation is based on *z*-transformed values). The average sample sizes of studies in the meta-analysis are shown by the size of circles and triangles (smaller circles and triangles represent smaller average sample sizes). Average correlations and confidence intervals were compared to the values in Table 1.


Frequency
30
20
10
0
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.22
0.24
0.26
0.28
0.30
0.32
Estimate of Population Standard Deviation


Hedges & Vevea Method
Hunter-Schmidt Method
$\bar{\rho} = .66$
$\bar{\rho} = .50$
$\bar{\rho} = .30$
$\bar{\rho} = .10$
$\bar{\rho} = .00$
Deviation From Population Effect Size
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
$\sigma_\rho = .04$
$\sigma_\rho = .08$
$\sigma_\rho = .16$
$\sigma_\rho = .32$
5 10 20 40 80160
Number of Studies in the Meta-Analysis


Mean Effect Size in Superpopulation = 0.0
Mean Effect Size in Superpopulation = 0.1
Mean Effect Size in Superpopulation = 0.3
Mean Effect Size in Superpopulation = 0.5
Mean Effect Size in Superpopulation = 0.8
Superpopulation Based on r
Superpopulation Based on z
Standard Deviation in Superpopulation = 0.04
Standard Deviation in Superpopulation = 0.08
Standard Deviation in Superpopulation = 0.16
Standard Deviation in Superpopulation = 0.32
Frequency
Effect Size in Population


Hedges & Vevea Method
Hunter-Schmidt Method
$\bar{\rho} = .80$
$\bar{\rho} = .50$
$\bar{\rho} = .30$
$\bar{\rho} = .10$
$\bar{\rho} = .00$
Deviation From Population Effect Size
$\sigma_\rho = .04$
$\sigma_\rho = .08$
$\sigma_\rho = .16$
$\sigma_\rho = .32$
5 10 20 40 80160
Number of Studies in the Meta-Analysis


Hedges & Vevea Method
Hunter-Schmidt Method
$\overline{\rho} = .66$
$\overline{\rho} = .50$
$\overline{\rho} = .30$
$\overline{\rho} = .10$
$\overline{\rho} = .00$
Deviation From Population Effect Size
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
$\sigma_\rho = .04$
$\sigma_\rho = .08$
$\sigma_\rho = .16$
$\sigma_\rho = .32$
5 10 20 40 80160
Number of Studies in the Meta-Analysis

FOOTNOTES

[i] These conversions can have statistical implications.

[ii] A mixed-effects model exists too in which population effect sizes differ but their variability is explained by a moderator variable that is treated as 'fixed' (see Overton, 1998) and also includes additional random heterogeneity.

[iii] Several studies provided multiple values because each study typically involved several meta-analyses conducted on different predictors or outcomes.

[iv] A third method developed by Rosenthal and Rubin (see Rosenthal, 1991) is popular but exists only in a fixed-effect form and differs from Hedges' method only in how the significance of the mean weighted effect size is calculated (see Field, 2001).

[v] This is the average effect size used in the fixed-effects model.

[vi] In both methods these different study weights consequently affect the estimates of the standard error: in the H-V method the standard error is clearly related to the study weights (see equation 10) and in the H-S method, the standard error (in equation 16) is based on the variance of observed correlations (equation 13), which is also a function of the study weights.

[vii] In this case the correlation standard deviation will be smaller than these values, because it was applied to a distribution of *z*s rather than *r*s (see Table 1).

[viii] I am grateful to Frank Schmidt for pointing this out.

[ix] A smoothing routine was used to plot the curves of the distributions and this resulted in the curves dropping below 0 in places and some other irregularities. Of

course frequencies were never below 0, but these curves give an overall impression of the shape of the distribution.

[x] These significance values need to be treated cautiously because these distributions are based on large samples and, as Field (2005) notes, this results in low standard errors and, therefore, large values of *z*.

[xi] To save space, the results for this simulation are not presented (but are available from the author).